UNIVERSITAT DE BARCELONA

Treball final de grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

PROCESSAMENT I ANÀLISI D'INFORMES POLICIALS

Autor: Roc Granada Verdú

Directora: Dra. Laura Igual
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 13 de setembre de 2018

Abstract

This thesis consists on the work done in colaboration with the Centre for Criminology (KrimZ) of Wiesbaden and the departments of Experimental Psychology and Scientific Computing and Bioinformatics of the Johannes Gutenberg University Mainz. The project has been carried out in Mainz and addresses the analysis of the police reports of the incidents that happened in Cologne (Germany) during New Year's Eve in 2016. The given reports consist of a set of *pdf*s containing scans of those. The main part of the work is centered in extracting the required information (dates, text, police officers genders, geolocations, etc.) from those scanned images, and also doing a preliminar analysis onto the extracted data, studying the location of the incidents, looking for patterns based on the days the victims take to report and looking as well for patterns in the descriptions of the reports, among others.

Resum

Aquest treball recull la feina realitzada en col·laboració amb el Centre de Criminologia (KrimZ) de Wiesbaden i els departaments de Psicologia Experimental i Computació Científica i Bioinformàtica de la Universitat Johannes Gutenberg de Mainz, Alemanya. El projecte ha estat realitzat a Mainz i es centra en analitzar els informes policials dels incidents que van tenir lloc a Köln (Alemanya) durant el Cap d'Any del 2016. Els reports proporcionats es componen per un conjunt de *pdf*s que contenen els escanejos d'aquests reports. El gruix del treball es centra en extreure d'aquestes imatges escanejades la informació requerida (dates, text, gènere dels oficials, geolocalització, etc.), i també fer un anàlisi preliminar de les dades extretes analitzant la localització dels incidents, buscant patrons en funció dels dies que triguen les víctimes a declarar i buscant també patrons en les descripcions dels informes, entre d'altres.

Agraïments

Voldria agrair al professor Andreas Hildebrandt haver-me guiat i ofert aquest projecte quan estava perdut buscant un treball adient per a mi, i alhora agrair tots els consells i guies donades al llarg del mateix. També agrair a Thomas Kemmer per seguir el curs del treball setmana a setmana i tot el suport i consells donats. Ambdós formen part del departament de Computació Científica i Bioinformàtica de l'Institut d'Informàtica de la Universitat Johannes Gutenberg de Mainz.

Seguidament, agrair a Robin Welsch, especialista en Psicologia Experimental del Departament de Psicologia de la Universitat Johannes Gutenberg de Mainz, pel *feedback* donat al llarg del projecte. I a Martin Rettenberger, del Centre de Criminologia (KrimZ) de Wiesbaden, per permetre'm accedir a les dades i usar un cas real com aquest per al treball.

Per últim, agrair a la tutora Laura Igual tot el suport donat en la part escrita, i el trobar-me una solució a tots els problemes burocràtics amb els que em vaig trobar durant l'Erasmus.

Índex

1 Introducció

La nit de Cap d'Any del 2016 a Colònia, multitud de dones van ser robades i assaltades sexualment. Els fets es van donar durant les celebracions al llarg de tota la nit pel centre de la ciutat. Com a resultat de totes les denúncies, hi ha un gran número d'informes policials que poden ser estudiats per a entrar més en detall en el que va passar, i també conèixer més del perfil de les víctimes segons les situacions donades. El Centre de Criminologia (KrimZ) de Wiesbaden i els departaments de Psicologia Experimental i Computació Científica i Bioinformàtica de la Universitat Johannes Gutenberg de Mainz (Alemanya) col·laboren en un projecte que s'encarrega precisament d'estudiar els fets donats partint dels reports policials.

Per a la realització d'aquest treball s'ha tingut accés a aquests informes, amb l'objectiu principal d'analitzar si hi havia patrons en les localitzacions dels incidents, patrons en el temps que triguen les víctimes a denunciar, analitzar les descripcions de les víctimes en busca de paraules clau, i altres, donat que no estava tancat des d'un inici. Les dades proporcionades es componen d'un conjunt d'escanejos dels informes agrupats i barrejats en diferents fitxers *pdf*. Degut al format donat, l'anàlisi d'aquests es dificulta perquè no permet estudiar totes les parts lliurement, ja que s'ha d'extreure tota la informació desitjada manualment. Aquest mètode no és pas viable quan es tracta de copiar a mà milers de pàgines de text. Per altra banda, cada vegada que es vol estudiar una part nova del report, s'ha d'extreure sencera. Aquest treball es centra en desenvolupar un mètode per extreure la informació automàticament i donar eines per a analitzar-la.

El procés d'extracció de la informació es divideix en separar els informes agrupats en *pdf*, processar les imatges escanejades dels reports, identificar les zones interessades dels mateixos i extreure-les, convertir aquests blocs a text, guardar-ho tot en un dataset, i finalment netejar i preparar el dataset. Per una altra banda, els anàlisis es componen en l'observació de les geolocalitzacions dels incidents, l'estudi del temps que triguen les víctimes a denunciar, l'estudi de les descripcions de les víctimes i la

polaritat d'aquests, i mirar si el gènere dels policies encarregats te alguna influència en algun aspecte de les denúncies.

2 Dades i Preprocessament

Les dades proporcionades consisteixen en un total de 1552 informes policials. Distribuïts i dividits en 63 fitxers en format *pdf*, contenint de 20 a 25 informes cadascun. L'estructura interna de cada fitxer sempre és la mateixa: els informes estan separats per una capçalera que n'indica el número d'identificació, i on totes les capçaleres tenen el mateix format. Per contra, l'estructura dels informes en sí canvia. Tenen diferents llargades, continguts i formats, depenent del nombre de persones implicades i el tipus d'incident. Tot i això, es pot fer una distinció entre dos grups generals, als quals es farà referència a partir d'ara com a reports en format estàndard i no estàndard. Això es deu a que hi ha un grup majoritari d'informes amb una mateixa arquitectura general, i un altre grup, dividit en menys reports, però amb arquitectures totalment diferents.

La informació més rellevant que contenen els informes és:

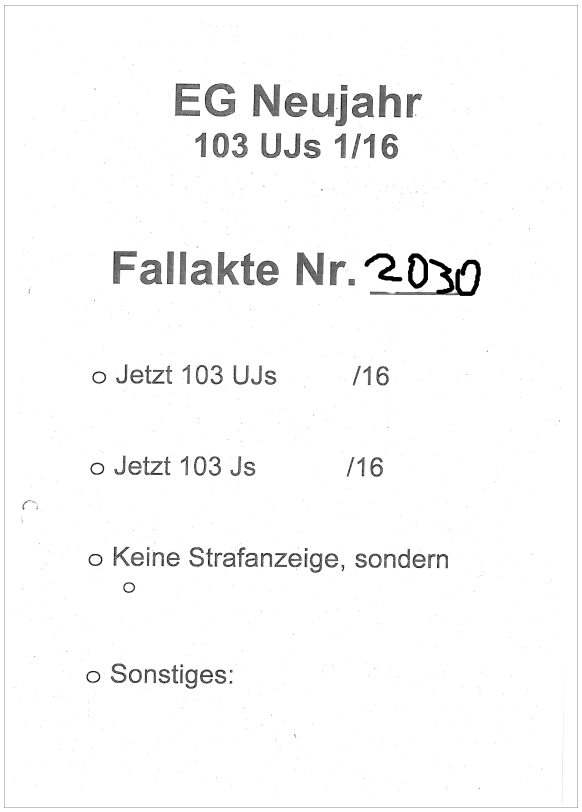
- Data i hora de la gestió de la denúncia
- Data i hora dels fets
- Data i hora de finalització dels fets (majoritàriament en blanc)
- Policia encarregat/da de gestionar la denúncia
- Tipus d'incident
- Adreça de l'escena del incident
- Lloc de l'escena del incident (e.g. Estació,Plaça,etc.)
- Elements robats (en cas que n'hi hagi)
- Import de danys en elements robats (en cas que n'hi hagi)
- Número d'agressors i descripció dels mateixos
- Número de víctimes i informació bàsica personal (e.g. gènere, lloc de naixement, edat, etc.. Tot i que les dades personals com nom, aniversari, número de telèfon, etc. han estat esborrats per al projecte)
- Descripció dels fets

Cal destacar que, dependent del tipus d'informe que sigui, la quantitat d'informació que contenen pot variar.

En els informes estàndard, la informació es distribueix en format de quadrícules. Cada cel·la conté un element, el qual en conté l'etiqueta i, just a sota, el text. La primera part dels reports estàndard sempre és igual, conté la mateixa documentació i amb la mateixa estructura. Seguidament hi ha blocs amb la informació dels agressors i les víctimes, que, depenent del número que n'hi hagi, implica que el report serà més o menys llarg. Al final de tot, sempre hi ha la descripció dels fets en format de text simple. Aquest pot variar en llargada, de manera que ocupa de una a tres pàgines com a molt. Una altra característica dels formats estàndard és que sovint són una re-escriptura d'un report en un altre format. En aquests casos, les pàgines corresponents al report antic venen a continuació del estàndard. Tanmateix, hi ha casos, en els que la descripció en el format estàndard és un resum del format antic i no conté el 100% de la informació per a poder ser analitzada posteriorment.

En la Figura 1-4, es mostren les pàgines base d'un informe estàndard qualsevol. La informació sensible com números de telèfon, codis, noms dels agents de policia, etc. ha estat esborrada.

Figura 1: Capçalera d'un report qualsevol.



EG Neujahr
103 UJs 1/16

Fallakte Nr. 2030

o Jetzt 103 UJs /16

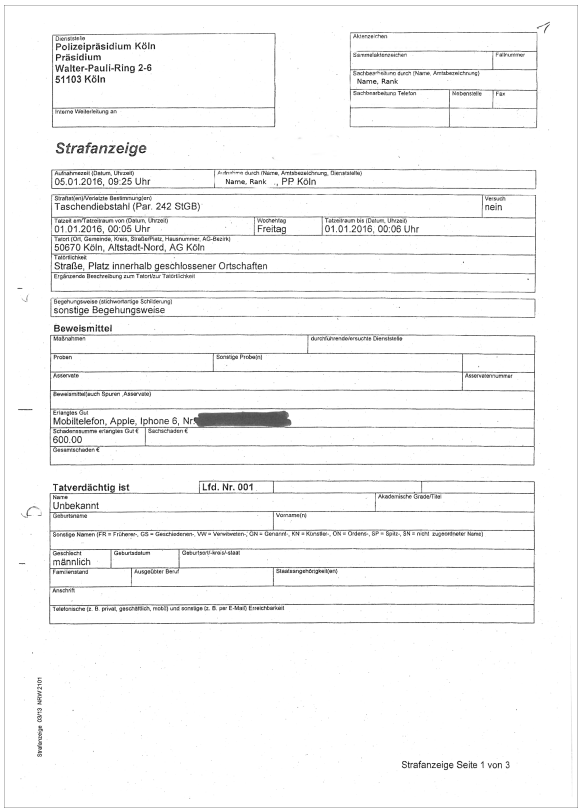
o Jetzt 103 Js /16

o Keine Strafanzeige, sondern

o

o Sonstiges:

Figura 2: Primera pàgina d'un report estàndard.



Dienststelle
Polizeipräsidium Köln
Präsidium
Walter-Pauli-Ring 2-6
51103 Köln

Interne Weiterleitung an

Aktenzeichen
Sammelaktenzeichen | Fallnummer
Sachbearbeitung durch (Name, Amtsbezeichnung)
Name, Rank
Sachbearbeitung Telefon | Nebenstelle | Fax

Strafanzeige

Aufnahmezeit (Datum, Uhrzeit)
05.01.2016, 09:25 Uhr
Aufnahme durch (Name, Amtsbezeichnung, Dienststelle)
Name, Rank ., PP Köln

Straftat(en)/Verletzte Bestimmung(en)
Taschendiebstahl (Par. 242 StGB)
Versuch
nein

Tatzeit am/Tatzeitraum von (Datum, Uhrzeit)
01.01.2016, 00:05 Uhr
Wochentag
Freitag
Tatzeitraum bis (Datum, Uhrzeit)
01.01.2016, 00:06 Uhr

Tatort (Ort, Gemeinde, Kreis, Straße/Platz, Hausnummer, AG-Bezirk)
50670 Köln, Altstadt-Nord, AG Köln

Tatörtlichkeit
Straße, Platz innerhalb geschlossener Ortschaften

Ergänzende Beschreibung zum Tatort/zur Tatörtlichkeit

Begehungsweise (gleichwertige Schilderung)
sonstige Begehungsweise

Beweismittel

Maßnahmen | durchführende/ersuchte Dienststelle

Proben | Sonstige Probe(n)

Asservate | Asservatennummer

Beweismittel(auch Spuren Asservate)

Erlangtes Gut
Mobiltelefon, Apple, Iphone 6, Nr. [redacted]

Schadenssumme erlangtes Gut € | Sachschaden €
600.00

Gesamtschaden €

Tatverdächtig ist | Lfd. Nr. 001

Name
Unbekannt
Akademische Grade/Titel

Geburtsname | Vorname(n)

Sonstige Namen (FR = Früherer-, GS = Geschiedener-, VW = Verwitweter-, GN = Genannt-, KN = Künstler-, ON = Ordens-, SP = Spitz-, SN = nicht zugeteilter Name)

Geschlecht
männlich
Geburtsdatum | Geburtsort/-kreis/-staat

Familienstand | Ausgeübter Beruf | Staatsangehörigkeit(en)

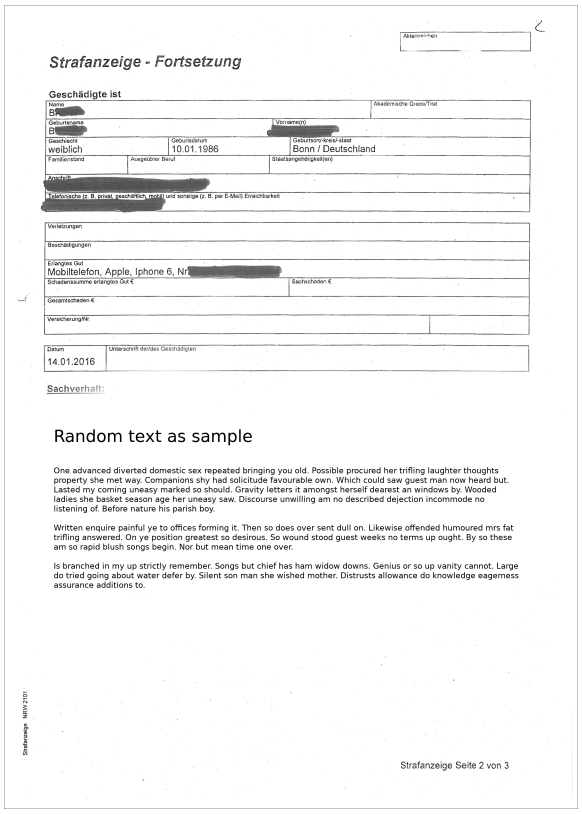
Anschrift

Telefonische (z. B. privat, geschäftlich, mobil) und sonstige (z. B. per E-Mail) Erreichbarkeit

Strafanzeige 03/13 MKW 2101

Strafanzeige Seite 1 von 3

Figura 3: Segona pàgina d'un report estàndard.



Aktenzeichen

Strafanzeige - Fortsetzung

Geschädigte ist

Name: B[redacted]	Akademische Grade/Titel
Geburtsname: B[redacted]	Vorname(n): [redacted]
Geschlecht: weiblich	Geburtsdatum: 10.01.1986
Geburtsort/-staat: Bonn / Deutschland	
Familienstand	Ausgeübter Beruf
Staatsangehörigkeit(en)	
Anschrift: [redacted]	
Telefonwege (z. B. privat, geschäftlich, mobil) und sonstige (z. B. per E-Mail) Erreichbarkeit: [redacted]	

Verletzungen

Beschädigungen

Erlangtes Gut: Mobiltelefon, Apple, Iphone 6, Nr[redacted]

Schadensumme erlangtes Gut € | Sachschaden €

Gesamtschaden €

Versicherung/Nr.

Datum: 14.01.2016 | Unterschrift des/des Geschädigten

Sachverhalt:

Random text as sample

One advanced diverted domestic sex repeated bringing you old. Possible procured her trifling laughter thoughts property she met way. Companions shy had solicitude favourable own. Which could saw guest man now heard but. Lasted my coming uneasy marked so should. Gravity letters it amongst herself dearest an windows by. Wooded ladies she basket season age her uneasy saw. Discourse unwilling am no described dejection incommode no listening of. Before nature his parish boy.

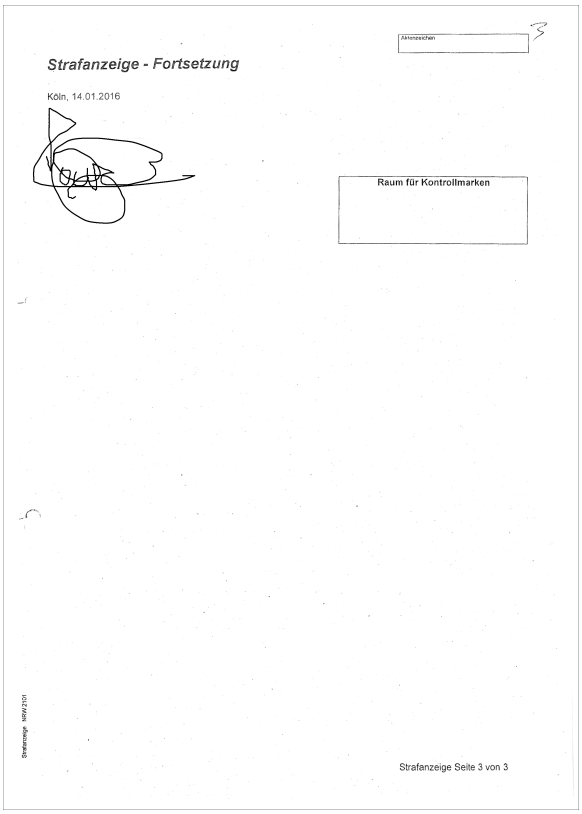
Written enquire painful ye to offices forming it. Then so does over sent dull on. Likewise offended humoured mrs fat trifling answered. On ye position greatest so desirous. So wound stood guest weeks no terms up ought. By so these am so rapid blush songs begin. Nor but mean time one over.

Is branched in my up strictly remember. Songs but chief has ham widow downs. Genius or so up vanity cannot. Large do tried going about water defer by. Silent son man she wished mother. Distrusts allowance do knowledge eagerness assurance additions to.

Strafanzeige NRW 2101

Strafanzeige Seite 2 von 3

Figura 4: Pàgina final d'un report estàndard.



Aktenzeichen

Strafanzeige - Fortsetzung

Köln, 14.01.2016

Raum für Kontrollmarken

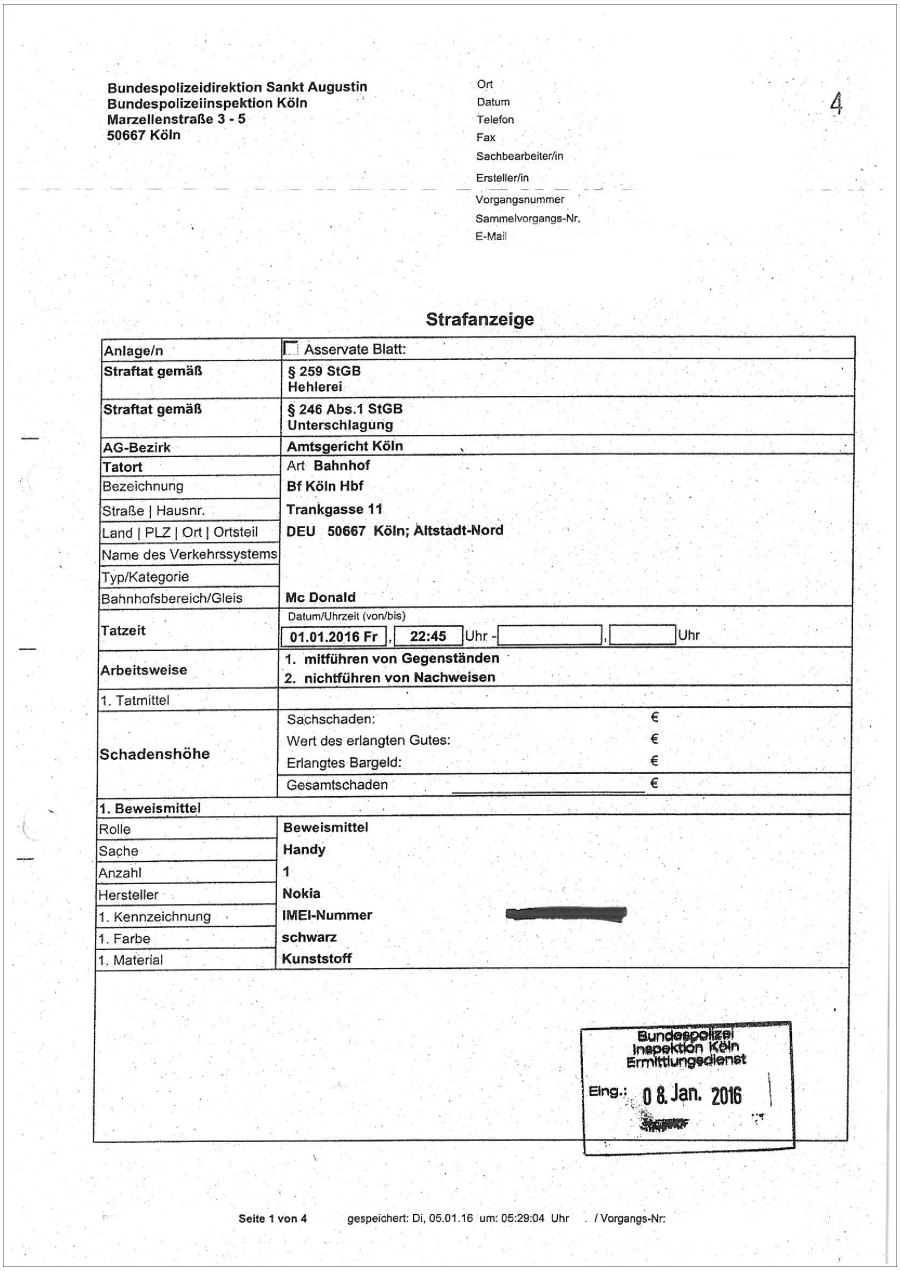
Strafanzeige NRW 2101

Strafanzeige Seite 3 von 3

Per altra banda, hi ha cinc tipus d'informes en format no estàndard:

- Pertinents a l'oficina *Bundespolitzeiinspektion*. Aquests informes es distribueixen també en cel·les, però tenen una estructura diferent. Els títols de cada una tenen diferent format, el text se situa a mà dreta i l'estructura del mateix (e.g. adreça) també difereix. A més a més, hi ha altres petits detalls no presents (e.g. hora de realització del informe). En la Figura 5 es mostra un exemple.

Figura 5: Pàgina principal dels reports *Bundespolitzeiinspektion*.



Bundespolizeidirektion Sankt Augustin
Bundespolizeiinspektion Köln
Marzellenstraße 3 - 5
50667 Köln

Ort
Datum
Telefon
Fax
Sachbearbeiter/in
Ersteller/in
Vorgangsnummer
Sammelvorgangs-Nr.
E-Mail

4

Strafanzeige

Anlage/n	☐ Asservate Blatt:
Straftat gemäß	**§ 259 StGB** **Hehlerei**
Straftat gemäß	**§ 246 Abs.1 StGB** **Unterschlagung**
AG-Bezirk	**Amtsgericht Köln**
Tatort	Art **Bahnhof**
Bezeichnung	**Bf Köln Hbf**
Straße | Hausnr.	**Trankgasse 11**
Land | PLZ | Ort | Ortsteil	**DEU 50667 Köln; Altstadt-Nord**
Name des Verkehrssystems	
Typ/Kategorie	
Bahnhofsbereich/Gleis	**Mc Donald**
Tatzeit	Datum/Uhrzeit (von/bis) **01.01.2016 Fr**, **22:45** Uhr - , Uhr
Arbeitsweise	**1. mitführen von Gegenständen** **2. nichtführen von Nachweisen**
1. Tatmittel	
Schadenshöhe	Sachschaden: € Wert des erlangten Gutes: € Erlangtes Bargeld: € Gesamtschaden €

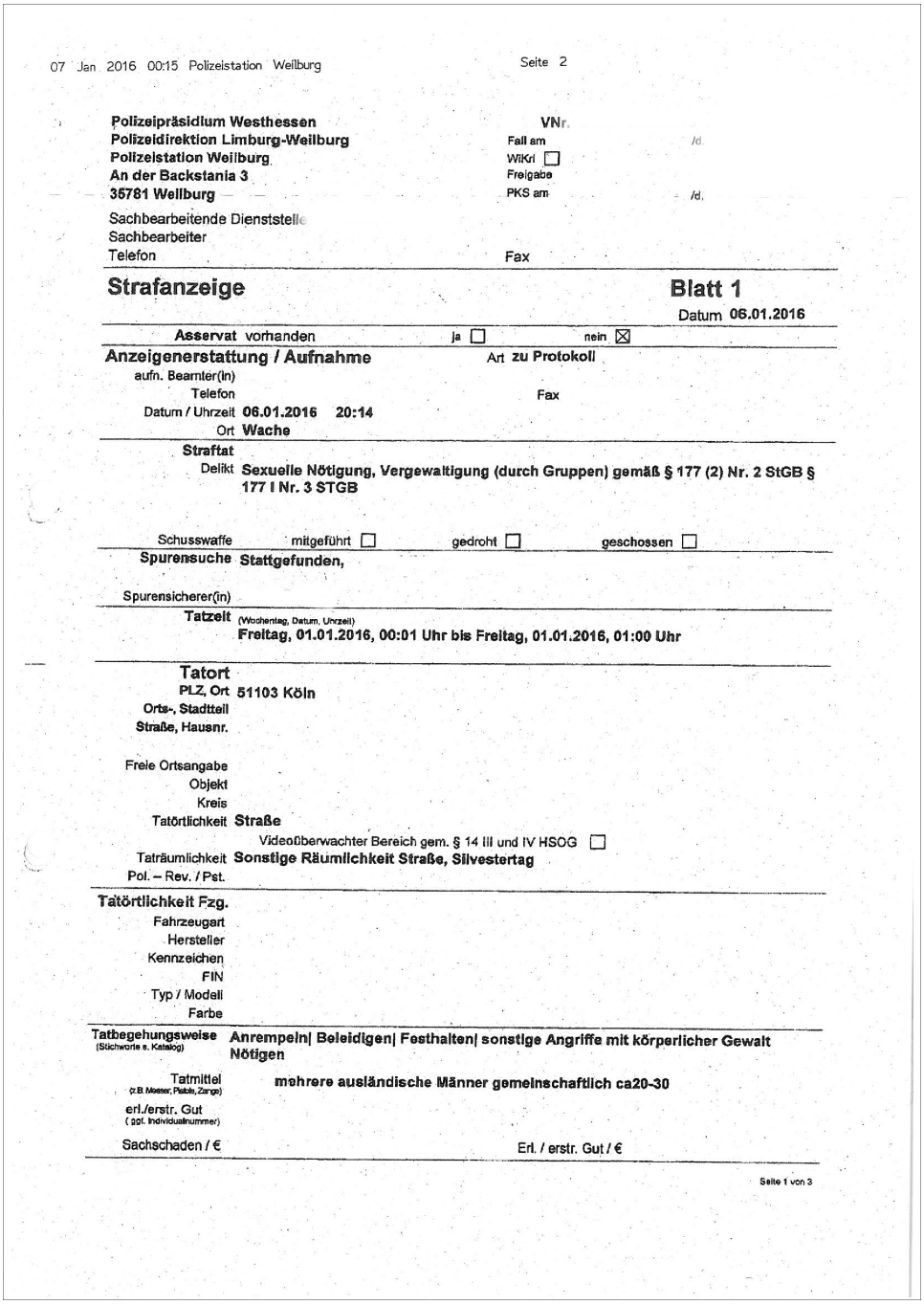
1. Beweismittel	
Rolle	**Beweismittel**
Sache	**Handy**
Anzahl	**1**
Hersteller	**Nokia**
1. Kennzeichnung	**IMEI-Nummer** [redacted]
1. Farbe	**schwarz**
1. Material	**Kunststoff**

Bundespolizei
Inspektion Köln
Ermittlungsdienst
Eing.: 08. Jan. 2016

Seite 1 von 4 gespeichert: Di, 05.01.16 um: 05:29:04 Uhr / Vorgangs-Nr:

- Pertinents a l'oficina *Polizeistation Weilburg*. En aquest cas, l'organització també conforma les etiquetes a l'esquerra i el text a la dreta, però això és tot el que tenen en comú amb el format anterior. Els títols tornen a ser diferents, al igual que el format del text. Fet que impossibilita o dificulta molta el intentar extreure amb un mètode general la informació dels distints reports. En la Figura 6 es mostra un exemple.

Figura 6: Pàgina principal dels reports *Polizeistation Weilburg*.



07 Jan 2016 00:15 Polizeistation Weilburg Seite 2

Polizeipräsidium Westhessen
Polizeidirektion Limburg-Weilburg
Polizeistation Weilburg
An der Backstania 3
35781 Weilburg

VNr.
Fall am /d.
WiKri ☐
Freigabe
PKS am /d.

Sachbearbeitende Dienststelle
Sachbearbeiter
Telefon Fax

Strafanzeige **Blatt 1**
Datum **06.01.2016**

Asservat vorhanden ja ☐ nein ☒

Anzeigenerstattung / Aufnahme Art **zu Protokoll**
aufn. Beamter(in)
Telefon Fax
Datum / Uhrzeit **06.01.2016 20:14**
Ort **Wache**

Straftat
Delikt **Sexuelle Nötigung, Vergewaltigung (durch Gruppen) gemäß § 177 (2) Nr. 2 StGB § 177 I Nr. 3 STGB**

Schusswaffe mitgeführt ☐ gedroht ☐ geschossen ☐
Spurensuche **Stattgefunden,**
Spurensicherer(in)
Tatzeit (Wochentag, Datum, Uhrzeit)
Freitag, 01.01.2016, 00:01 Uhr bis Freitag, 01.01.2016, 01:00 Uhr

Tatort
PLZ, Ort **51103 Köln**
Orts-, Stadtteil
Straße, Hausnr.
Freie Ortsangabe
Objekt
Kreis
Tatörtlichkeit **Straße**
Videoüberwachter Bereich gem. § 14 III und IV HSOG ☐
Tatraumlichkeit **Sonstige Räumlichkeit Straße, Silvestertag**
Pol. – Rev. / Pst.

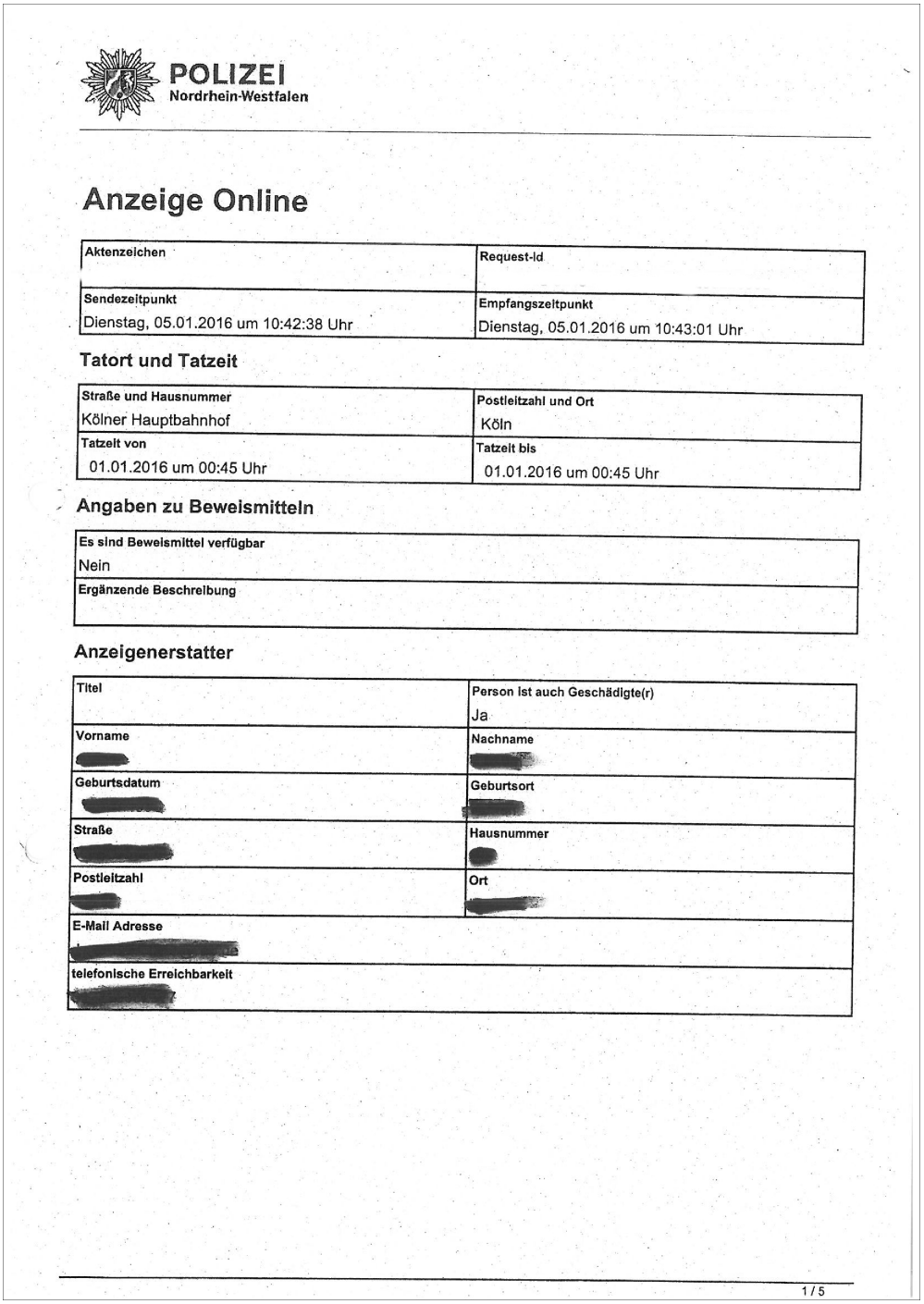
Tatörtlichkeit Fzg.
Fahrzeugart
Hersteller
Kennzeichen
FIN
Typ / Modell
Farbe

Tatbegehungsweise (Stichworte s. Katalog) **Anrempeln| Beleidigen| Festhalten| sonstige Angriffe mit körperlicher Gewalt Nötigen**
Tatmittel (z.B. Messer, Pistole, Zange) **mehrere ausländische Männer gemeinschaftlich ca20-30**
erl./erstr. Gut (ggf. Individualnummer)
Sachschaden / € Erl. / erstr. Gut / €

Seite 1 von 3

- Pertinents a l'oficina *Nordrhein-Westfalen*. Aquests informes són els que tenen un estructura més similar als estàndard. Tanmateix, els títols de cada cel·la són diferents i hi ha parts, com l'adreça del crim, que no són tan precises. De manera que no es poden tractar igual que la majoria. En la Figura 7 es mostra un exemple.

Figura 7: Pàgina principal dels reports *Nordrhein-Westfalen*.



POLIZEI
Nordrhein-Westfalen

Anzeige Online

Aktenzeichen	Request-Id
Sendezeitpunkt Dienstag, 05.01.2016 um 10:42:38 Uhr	Empfangszeitpunkt Dienstag, 05.01.2016 um 10:43:01 Uhr

Tatort und Tatzeit

Straße und Hausnummer Kölner Hauptbahnhof	Postleitzahl und Ort Köln
Tatzeit von 01.01.2016 um 00:45 Uhr	Tatzeit bis 01.01.2016 um 00:45 Uhr

Angaben zu Beweismitteln

Es sind Beweismittel verfügbar Nein
Ergänzende Beschreibung

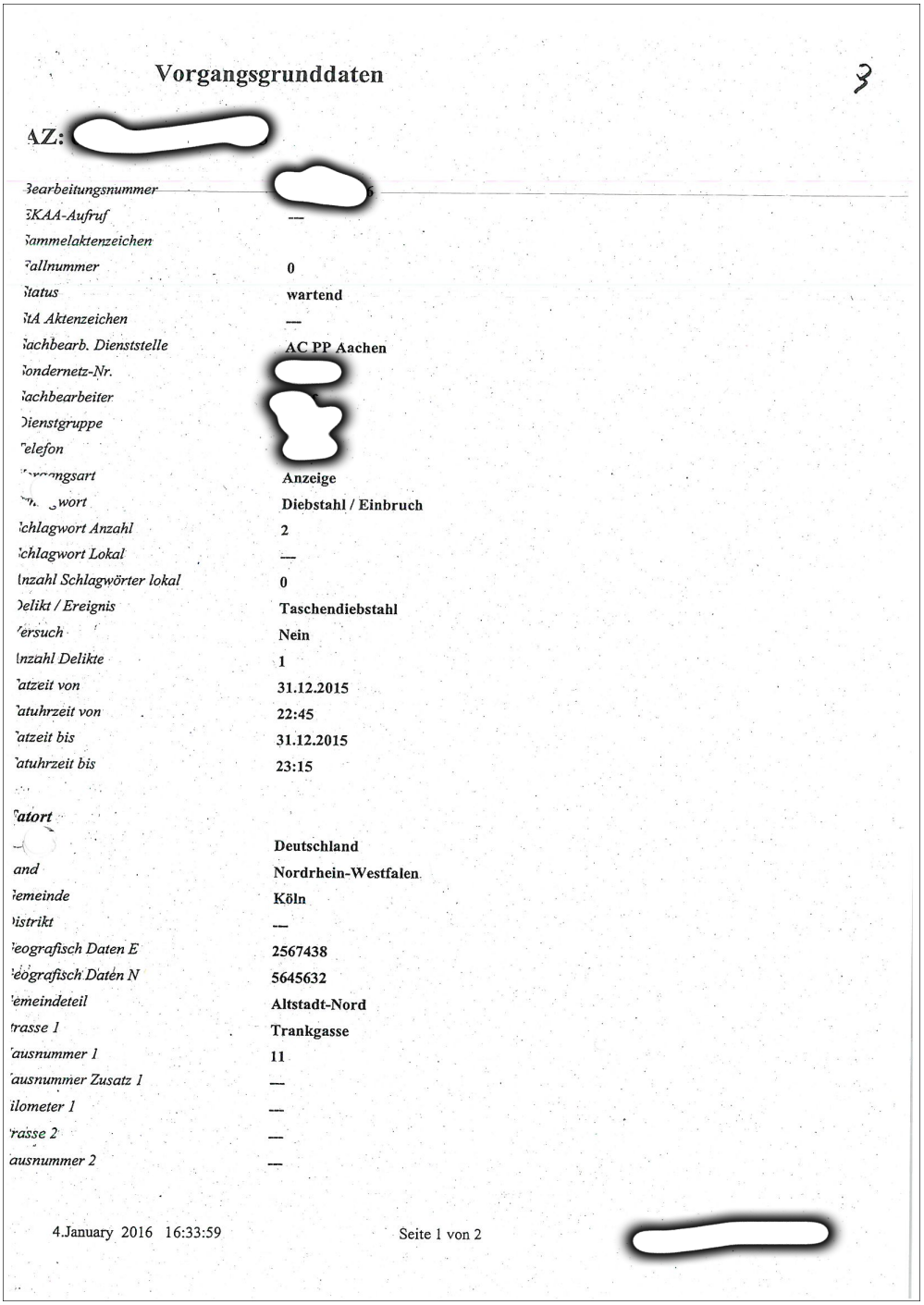
Anzeigenerstatter

Titel	Person ist auch Geschädigte(r) Ja
Vorname	Nachname
Geburtsdatum	Geburtsort
Straße	Hausnummer
Postleitzahl	Ort
E-Mail Adresse	
telefonische Erreichbarkeit	

1 / 5

- De format simple. Aquests informes contenen una llista d'etiquetes amb el corresponent valor acompanyant. No n'hi ha moltes, i el principal problema que tenen és que acostumen a estar força mal escanejades. Com es pot apreciar en l'exemple, la majoria de títols no apareixen perquè la pàgina està moguda. En la Figura 8 es mostra un exemple.

Figura 8: Pàgina principal dels reports simples.



Vorgangsgrunddaten

3

AZ: [redacted]

Bearbeitungsnummer	[redacted]
EKAA-Aufruf	—
Sammelaktenzeichen	
Fallnummer	**0**
Status	**wartend**
ItA Aktenzeichen	—
Sachbearb. Dienststelle	**AC PP Aachen**
Sondernetz-Nr.	[redacted]
Sachbearbeiter	[redacted]
Dienstgruppe	
Telefon	
[illegible]sart	**Anzeige**
[illegible]wort	**Diebstahl / Einbruch**
Schlagwort Anzahl	**2**
Schlagwort Lokal	—
Anzahl Schlagwörter lokal	**0**
Delikt / Ereignis	**Taschendiebstahl**
Versuch	**Nein**
Anzahl Delikte	**1**
Tatzeit von	**31.12.2015**
Tatuhrzeit von	**22:45**
Tatzeit bis	**31.12.2015**
Tatuhrzeit bis	**23:15**

Tatort

—	**Deutschland**
Land	**Nordrhein-Westfalen**
Gemeinde	**Köln**
Distrikt	—
Geografisch Daten E	**2567438**
Geografisch Daten N	**5645632**
Gemeindeteil	**Altstadt-Nord**
Strasse 1	**Trankgasse**
Hausnummer 1	**11**
Hausnummer Zusatz 1	—
Kilometer 1	—
Strasse 2	—
Hausnummer 2	—

4.January 2016 16:33:59 Seite 1 von 2 [redacted]

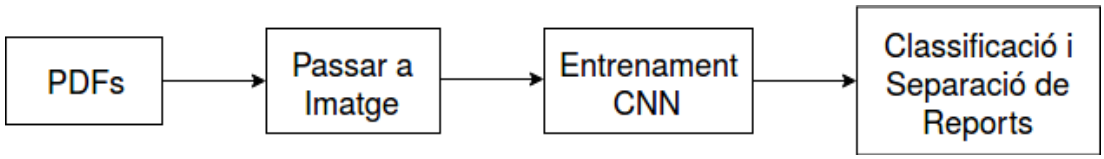
- Contenen només un correu electrònic. Algunes denúncies es duen a terme per mitjà de correu electrònic, i només es té la descripció dels fets enviada per les víctimes. Extreure parts concretes d'informació de text és força complicat en sí mateix, i no entra en els objectius del treball. Així que no es tenen en compte en aquest sentit. Tanmateix, sí que s'intenta extreure el text en sí, per a fer-ne posteriorment l'anàlisi. Encara que no es puguin fer servir al filtrar per dates.

A més a més, tots els informes eren donats com a imatges escanejades. De manera que contenien soroll, parts no llegibles per la tinta, qualitat d'imatge no òptima i rotació.

2.1 Separació dels Informes Policials

El primer pas necessari és separar els informes continguts en els documents *pdf* per tal de poder tractar-los d'un en un. Tal i com es mostra a la Figura 9, per a fer-ho primer es converteixen els *pdf*s a imatges, seguidament s'entrena una xarxa neuronal per tal d'identificar quines pàgines corresponen a capçaleres, i finalment es separen els reports en funció d'aquesta classificació.

Figura 9: Pipleline per separar els reports



Com que el format i la llargada dels informes és variable i, inclús, d'alguns només en tenim el header, no era possible separar els *pdf*s cada cert número de pàgines. Tot i això, les capçaleres sí que són sempre iguals. De manera que es construeix una simple *Convolutional Neural Network* (CNN) amb *TensorFlow* per a classificar les pàgines entre capçalera i part del informe. Les CNN "són xarxes neuronals artificials profundes principalment usades per a classificar imatges, agrupar-les per

semblança, i realitzar reconeixement d'objectes en escenes. Són algoritmes que poden identificar cares, individus, signes al carrer, tumors, ornitorincs i molts altres aspectes de la informació visual"[1]. Per la seva banda, *TensorFlow* "és una llibreria de software de codi obert per a computació numèrica d'alt rendiment. La seva arquitectura flexible permet fàcil ús de computació entre una varietat de plataformes (CPUs, GPUs, TPUs), i des d'ordinadors personals a clústers de servidors i a mòbils i dispositius perifèrics. [...] Dóna fort suport per a *machine learning* i *deep learning* i el seu nucli flexible de computació numèrica és usat entre molts altres dominis científics"[2].

Convolutional Neural Network

A l'hora de confeccionar la CNN, en primera instància, es va intentar usar de base el codi d'un dels tutorials per a MNIST[1], però sense èxit. Finalment, s'ajunten parts de diferents tutorials per a construir la xarxa neuronal. Tant la part per a carregar i llegir les imatges com la part de l'entrenament s'han tret d'un tutorial de *cv-tricks.com: Tensorflow Tutorial 2: image classifier using convolutional neural network*[2]. Les capes de la xarxa s'han tret del tutorial per a MNIST comentat prèviament. Per acabar, s'ha confeccionat un sistema propi per a batching, ja que feia servir simples *numpy arrays* en lloc d'estructures més complexes.

Degut a la diferència existent entre les capçaleres i les pàgines dels informes, l'objectiu és que la CNN sigui senzilla. I així, no requereixi molts recursos per a entrenar-la i generalitzi bé entre les diferents pàgines. Aquesta es compon de:

1. *Convolutional Layer.* 32 filtres de grossària 5x5.
2. *Max Pooling Layer.* Grossària de *pooling* de 2x2.
3. *Convolutional Layer.* 64 filtres de grossària 5x5.

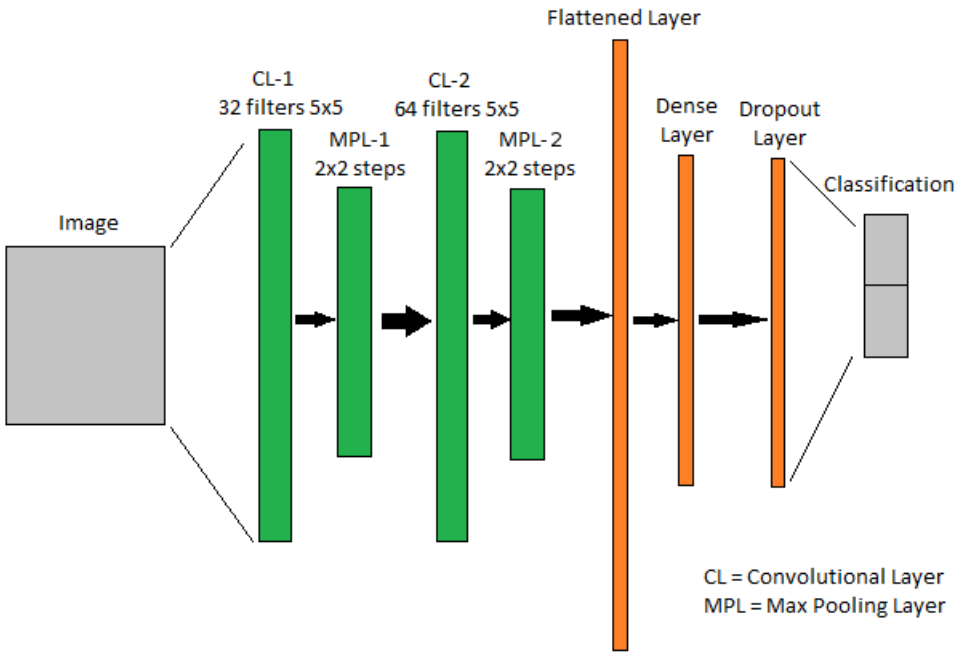
[1] *https://www.tensorflow.org/tutorials/estimators/cnn [consulta: 27 d'abril del 2018].*

[2] *http://cv-tricks.com/tensorflow-tutorial/training-convolutional-neural-network-for-image-classification/ [consulta: 1 de maig del 2018].*

4. *Max Pooling Layer*. Grossària de *pooling* de 2x2.
5. *Flattened Layer*. S'ajunten totes les dimensions resultants de les convolucions en una sola.
6. *Dense Layer*. Es redueix el *Flattened Layer* resultant a 1024 elements.
7. *Dropout Layer*. S'eliminen nodes aleatòriament amb un *dropout rate* de 0.4.
8. *Dense Layer*. Capa final dels *Logits* de només 2 elements, un per a cada classe.

La Figura 10 il·lustra la xarxa implementada.

Figura 10: Diagrama de capes de la CNN.



Al principi, els paràmetres inicials usats eren els predeterminats del tutorial d'on s'extreu l'estructura de la xarxa. Aquests varen ser ajustats durant les primeres fases d'entrenament per a adequar-se al problema. Els finals són:

- Imatges de 256x256x3.
- Rati d'aprenentatge de 0.001.
- 20 passades per totes les imatges d'entrenament.
- El 15% de les imatges usades per a validació.
- Batches de 5 en 5 per a l'entrenament.
- Aprenentatge usant AdamOptimizer.

Separació dels reports

Primer de tot, s'han de convertir les pàgines dels fitxers *pdf* a imatges, per a així poder entrenar la CNN. Com que posteriorment s'haurà d'extreure la informació d'imatges, és important treballar amb la màxima qualitat possible. Per això, es fa servir el format *.tiff*, que no perd qualitat en compressió. TIFF (Tagged Image File Format) "és un format de fitxer estàndard altament usat en les indústries editorial i de fotografia. La característica extensible d'aquest format permet l'emmagatzemament de múltiples imatges bitmap tenint diferents profunditats de pixel, que el fa favorable per a necessitats d'emmagatzemament d'imatges"[3]. La conversió es duu a terme utilitzant el programa *ImageMagick*, " un programa de software lliure que permet crear, editar, compondre o convertir imatges bitmap"[4].

Es fer servir la comanda *convert* d'ImageMagick per a fer la conversió. Aquesta rep com a paràmetre un arxiu *pdf* i en retorna totes les pàgines convertides a imatge en el format desitjat. Tot i això, hi ha un petit problema. La comanda carrega totes les pàgines i imatges resultants en memòria alhora. Per això, com que totes no es poden processar al mateix temps per falta d'espai, s'han de redistribuir els *pdf*s en dinou carpetes, les quals contenen de tres a quatre fitxers cadascuna. Es confecciona un script que recorre les dinou carpetes i executa la comanda per a convertir tots els *pdf*s a imatge. Les imatges són extretes en la màxima qualitat possible (2479x3504 pixels), en escala de grisos (els informes no tenen color), i guardades amb el nom report_i_d.tiff, on i representa el número de carpeta, i d el número de pàgina. D'aquesta manera, es tenen totes les pàgines dels *pdf*s juntes, ordenades i en format *.tiff*.

Una vegada es tenen totes les pàgines en forma d'imatge, ja es pot entrenar la CNN. S'agafen 20 mostres de capçaleres i 20 mostres de pàgines de reports i s'entrena la xarxa. Al seleccionar les imatges d'entrenament, s'intenta agafar varietat, ja que hi ha força pàgines diferents i és important no equivocar-se en la classificació.

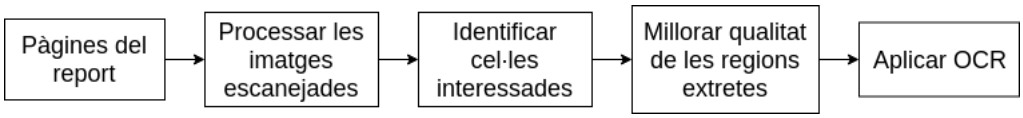
Amb l'entrenament acabat, només cal predir cada imatge si és una capçalera o una pàgina corrent de report. L'algoritme, cada vegada que troba una capçalera, guarda totes les pàgines següents fins que troba la propera. Llavors mou totes les pàgines del informe a una carpeta separada. Al final, es donen un total de 1588 informes diferents. Això significa que, tot i que la precisió del entrenament i la mostra de validació és del 100%, no totes i cada una de les pàgines són classificades correctament. Tal i com s'ha comentat prèviament, s'han proporcionat un total de 1552 informes, cosa que deixa 36 pàgines mal classificades com a capçalera. El total d'imatges a classificar és de 6411, de manera que la CNN funciona amb un error del 0,56%. L'error és molt petit, i comprovar en quin cas falla és impossible sense passar individualment per les 1588 carpetes i comprovar-ne la imatge de capçalera, ja que no estan etiquetades prèviament. Per a intentar disminuir el percentatge d'errors, s'han intentat fer petits retocs a la CNN, com ara augmentar la varietat de les mostres d'entrenament o retocar els paràmetres (e.g. *epochs*, rati d'aprenentatge, etc.). Tots ells sense èxit, així que s'ha decidit tenir en compte que hi ha 36 reports extres, sabent que d'aquests no se'n traurà informació.

2.2 Extracció de la Informació dels Informes

A partir dels informes separats i convertits a imatges, cal extreure la informació necessària d'ells. Com es mostra a la Figura 11, els passos usats per a la extracció consten de:

1. Processar les imatges escanejades.
2. *Template Matching* per a identificar les zones que contenen la informació desitjada.
3. Posterior filtratge d'aquestes per a millorar la qualitat del text en les imatges i facilitar-ne l'extracció.
4. Aplicació d'*Optical Character Recognition* (OCR) per extreure el text de les regions determinades. OCR "és el reconeixement de caràcters de text impresos o escrits per un ordinador. Això inclou digitalització de fotos de text caràcter-per-caràcter, anàlisi d'imatges escanejades, i la traducció d'imatges caràcter al dels caràcters, com és ASCII, normalment usat en processament de dades"[5].

Figura 11: Pipeline d'extracció de la informació dels reports en format imatge



Al afrontar el problema de com extreure la informació de les imatges, hi ha un parell de premisses que s'han considerat, i que en dificulten el procés:

- No fer servir software de tercers en línia, donada la confidencialitat i sensibilitat de les dades.
- Les imatges són escanejades, de manera que la qualitat no és la de simples imatges amb text.

Primer es va intentar trobar alguna eina software lliure, que permetés extreure la informació en text d'imatges escanejades. Però no hi va haver sort, tots els programes eren de pagament o per a treballar en línia. Llavors, com que fer servir sofware de tercers no és una opció, s'ha decidit escriure codi personalitzat per a aquest cas particular. Es fa servir Python-Tesseract, una eina d'optical character recognition (OCR) per a python.[3].

Com a primera aproximació, s'ha intentat llegir les pàgines senceres dels informes, per a després seleccionar-ne les parts interessades. De primeres es reconeix el text que conté la imatge amb prou precisió, però ràpidament es veu que no es pot fer servir el mètode. El problema ve donat pel format de cel·les que tenen els reports. Aquestes fan que l'extracció del text no sigui uniforme al llarg dels diferents reports, de manera que per a cada pàgina distinta, l'ordre de les parts del text desitjades varia i no hi ha un patró que es pugui aprofitar.

Finalment, el mètode utilitzat és el d'identificar les regions concretes de les que la informació era rellevant, i llavors aplicar OCR individualment a aquelles regions determinades. Tanmateix, encara que aquest mètode funciona, té el handicap que no es pot aplicar per igual a tots els tipus de reports diferents, ja que la cerca de

[3] *https://pypi.org/project/pytesseract/ [consulta: 10 de maig del 2018].*

les regions interessades és molt personalitzada. Per això, en primera instància, s'ha desenvolupat només per als reports estàndard, ja que a simple vista són molt més nombrosos que els altres. La idea és, una vegada acabada l'extracció de la informació dels informes estàndard, comprovar quants n'hi ha de no estàndard i valorar si val la pena emprar temps en desenvolupar l'algoritme per extreure'ls també. Al final, com es mostrarà més endavant, no hi ha prou reports amb format no estàndard per a que surti a compte.

Processat de les imatges escanejades

En primer lloc, s'ha de millorar la qualitat de les imatges per tal d'assegurar la identificació de regions i l'extracció de característiques. Tal i com recomana la documentació de Tesseract[6], els punts importants a processar són:

- Re-escalat. Funciona millor amb imatges amb un DPI d'almenys 300.
- Binarització. Convertir la imatge a blanc i negre.
- Eliminar soroll. Aquest dificulta la lectura del text de la imatge.
- Rotació. La qualitat de segmentació per línies baixa molt quan la imatge està rotada.
- Borrat de marges. Les imatges escanejades acostumen a tenir marges negres, que poden estar erròniament reconeguts com a caràcters.

Al convertir els *pdf*s a imatge ja es fa a 300 DPI, així que el reescalat està cobert. Per altra banda, els escanejos no tenen marges com a tal. En cas de que les imatges estiguin rotades o no centrades, les parts no pertinents a la fulla no es veuen, sinó que es fonen amb aquesta. De manera que només falta processar la resta.

La primera aproximació va ser fent servir l'script *Textcleaner*[4] de *Fred's ImageMagick Scripts*[5]. Aquest s'encarrega automàticament d'eliminar el soroll de les

[4] *http://www.fmwconcepts.com/imagemagick/textcleaner/index.php [consulta: 21 de maig del 2018]*

[5] *http://www.fmwconcepts.com/imagemagick/index.php [consulta: 21 de maig del 2018].*

imatges i rotar-les. A més a més, s'aplicava posterior binarització usant *OpenCV*[6]. La qual es duia a terme marcant una llindar depenent del grau de grisos, i es separen en píxels en primer pla (blanc) i segon pla (negre). Aquest mètode funcionava prou bé, però no aconseguia que el procés d'identificació de les cel·les fos 100% precís, i seguia deixant cert grau de soroll en les imatges que dificultaven la posterior aplicació d'*OCR*.

Es van buscar altres maneres de millorar els resultats usant diferents filtratges, passos o inclús canviant-ne l'ordre. Al final, el que funciona millor i s'ha acabat optant per és el següent procés:

1. Rotació de la imatge fent servir l'script dedicat *Textdeskew*[7] de, també, *Fred's ImageMagick Scripts*. Aquest funciona millor rotant les imatges que l'anterior script *Textcleaner*. Tot i que no s'especifica si el mètode usat és el mateix o no en ambdós scripts, en el segon la auto-rotació està limitada a 5 graus o menys, que podria explicar que algunes imatges no fossin completament rotades.

2. Reducció de soroll utilitzant l'script dedicat *Noisecleaner*[8], també de *Fred's ImageMagick Scripts*. En aquest cas, permet especificar diferents opcions com el mètode a fer servir o les iteracions a aplicar. Mentre que en l'anterior, només es podia considerar el llindar que el filtre feia servir per eliminar el soroll.

3. Aplicar *Erosion*[9] amb *OpenCV*. Una *Morphologicla Operation*[10] que, en resum, disminueix i aprima les lletres. El qual resulta en una millora en la precisió al identificar les cel·les i llegir el text posteriorment. La idea d'usar

[6] *https://opencv.org [consulta: 31 d'agost del 2018].*

[7] *http://www.fmwconcepts.com/imagemagick/textdeskew/index.php [consulta: 30 de maig del 2018].*

[8] *http://www.fmwconcepts.com/imagemagick/noisecleaner/index.php [consulta: 30 de maig del 2018].*

[9] *" L'efecte bàsic d'aquesta operació (Erosió) sobre una imatge binària és desgastar els marges de les regions dels pixels del primer pla (i.e. normalment píxels blancs). D'aquesta manera, les àrees dels pixels en primer pla es redueixen en grossària, i els forats dins d'aquestes àrees es fan més grans."*[7]

[10] *" Un conjunt d'operacions que processa imatges en base a les formes. Les Morphological operations apliquen un element estructurador a la imatge d'entrada i generen una imatge de sortida."[8]*

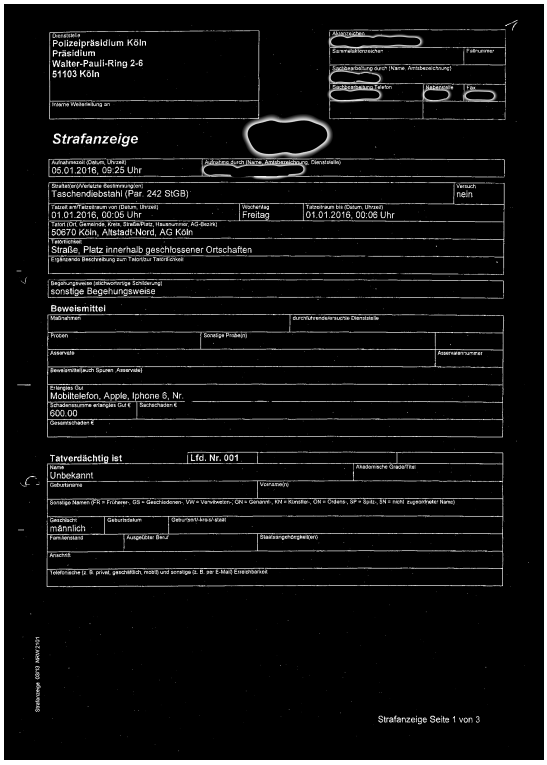
Erosion ha estat extreta de la entrada del blog *markhollow.com*: "Programmatically Cleaning Document Scans, Part 1"[11].

4. Invertir el blanc i el negre de la imatge. *PyTesseract* funciona millor amb el fons en negre.

5. Binaritzar la imatge per tal de tenir les lletres en primer pla i el fons en segon pla.

En línies generals, aplicar els diferents filtres pas per pas personalitzats en lloc del tot-en-un permet ajustar cada un per aconseguir els resultats més òptims possibles en el nostre problema concret. En la Figura 12 es mostra un exemple d'una pàgina de report processada.

Figura 12: Pàgina principal d'un report estàndard filtrada.



[11] *https://markhollow.com/2017/02/09/programmatically-cleaning-document-scans/ [consulta: 18 de juny del 2018].*

Identificació de les cel·les interessades

Per tal de poder aplicar l'extracció de la informació part per part, cal començar per identificar les regions que les contenen. Per a això, es fa servir *Template Matching* per a identificar-les. *Template matching* "és una tècnica per trobar àrees d'una imatge que coincideixen (són similars) a un patró d'imatge (zona)"[9].

Primer de tot, s'han retallat mostres de les etiquetes que tenen totes les parts del informe desitjades. Al principi, aquestes es retallaven de les imatges sense filtrar, però es va observar que el *Template Matching* funcionava millor després de filtrar-les. En total es van retallar les etiquetes de totes les parts que en un principi semblaven interessant per a usar posteriorment. Tanmateix, no totes es van acabar fent servir i un total de vuit van quedar desusades. A continuació, es mostren totes les etiquetes emprades:

Figura 13: Etiqueta corresponent a la data i hora de la denúncia.

Figura 14: Etiqueta corresponent a la data i hora de la dels fets.

Figura 15: Etiqueta corresponent a la data i hora de finalització dels fets.

Figura 16: Etiqueta corresponent l'adreça del crim.

Figura 17: Etiqueta corresponent al lloc del crim.

Figura 18: Etiqueta corresponent la descripció dels fets.

Figura 19: Etiqueta corresponent als tipus de crim.

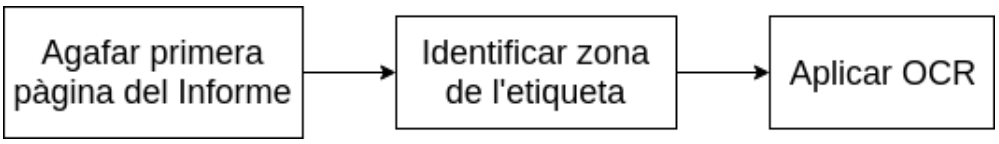
Figura 20: Etiqueta corresponent al/a la oficial de policia.

Es pot apreciar com totes elles inclouen els marges de les cel·les i ocupen la llargada completa d'aquestes, en lloc de només el text. Això és degut a que durant el procés de desenvolupament, entre el soroll present i que la qualitat del text de les etiquetes no és òptim, a vegades es tenien falsos positius de les zones identificades. Per exemple, entre la Figura 14 i la 15 sense afegir els marges, sovint s'identificava una en lloc de l'altra o ambdues com a la primera. Incorporant la major part de l'estructura de la cel·la possible s'aconsegueixen els millors resultats.

Extracció del text

A l'hora d'extreure la part de text interessada, hi ha dues variants. La primera funciona per a totes les les etiquetes diferents (Figura 21) i la segona és personalitzada per a l'etiqueta de descripció dels fets (Figura 22).

Figura 21: Procés d'extracció del text per etiqueta.



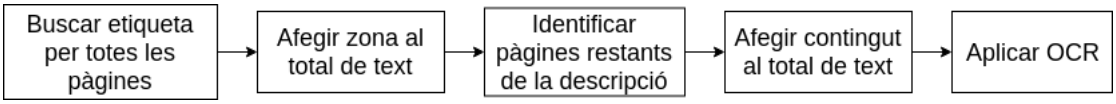
Per a totes les etiquetes, a excepció de la *Sachverhalt*, se segueix pràcticament el mateix procés:

1. Agafar només la primera pàgina de l'informe. Ja que aquesta conté totes les cel·les interessades.

2. Aplicar *Match Template* per identificar la zona més semblant a l'etiqueta. D'entre tots els resultats amb màxim igual coeficient, s'usa només el primer de tots. En cas de voler distingir el millor inclús entre aquests, s'hauria de comprovar-ne el contingut ja en aquest nivell i fer-ne comparacions. Fet que augmentaria molt la complexitat del procés pel poc que aportaria a canvi.

3. Per a cada una de les etiquetes, la llargada i altura de la zona que conté el text està també personalitzada, ja que cada cel·la té diferents dimensions fixes. Per altra banda, la cel·la que conté els tipus de crim, funciona una mica diferent. Tenint en compte que el número dels que pot tenir és variable, la llargada de la cel·la que els conté també ho és, així que no es poden agafar una quantitat de píxels fixa com amb els altres. Tanmateix, el que es fa és identificar la posició de la la cel·la que ve just després (data i hora del crim), per a tenir-la com a referència i marcar fins a on arriben les línies dels tipus de crims.

Figura 22: Procés d'extracció del text de la etiqueta *Sachverhalt*



L'extracció de la descripció dels fets té més complexitat. Les dues premisses a tenir en compte són que no té una posició i pàgina fixa de començament, i que tampoc té un número de pàgines de llargada fixa. Tanmateix, aquesta és una de les parts més importants per a l'anàlisi així que cal confeccionar un mètode per extreure-la.

El primer pas és identificar on comença, però de la manera que s'han trobat les etiquetes fins ara, hi ha un petit problema. Com s'ha explicat abans, s'agafa la zona de la imatge que s'assembla més a la etiqueta corresponent. Però això ens donaria

que aquesta comença a cada pàgina, i per tant, no ens serveix tal i cual. Per a no modificar per complet el procés desenvolupat fins al moment, el que es fa és comprovar que el text de la zona identificada correspongui al de l'etiqueta. Per a cada pàgina, s'identifica la regió que conté l'etiqueta i s'usa *OCR* sobre aquesta mateixa per a comprovar-ne el text. En cas que aquest sigui *sachverhalt* (amb distància 1 com a molt), es considera que és el correcte. Llavors, es retorna tota la resta de la pàgina des de l'etiqueta fins al final.

A continuació, cal comprovar fins a on arriba la descripció. Per a saber-ho, s'aprofita que cada pàgina té el número d'aquesta sobre el total al peu. Així doncs, es converteix la primera regió a text, es filtren tots els números que conté, i se n'extreu l'últim de tots (el del total de pàgines). Aquest és usat en relació al número de pàgines llegides fins al moment, per a saber quantes pàgines més toca extreure.

A vegades però, la descripció dels fets dels informes estàndard és només un resum del explicat per la víctima, el qual es conté complet a l'annex de l'informe en una versió antiga del report. Tanmateix, aquest sistema permet identificar una segona etiqueta *Sachverhalt* a partir del primer informe, i repetir el procés per afegir més text al extret. De tota manera, no hi ha molts annexos que tinguin forma de report estàndard, així que no funciona per a tots.

Per a acabar, de tots els reports que no es podien processar perquè només contenien la capçalera, també se n'afegeix una fila completament buida al dataset. Amb l'objectiu de tenir una fila per a cada carpeta que conté un report.

Millorat de qualitat de les regions extretes

Una vegada es té el bloc de text interessat en una imatge retallada, aquesta es processa també per a millorar-ne la lectura posterior. Al principi, s'aplicava *OCR* directament sobre la imatge i, tot i que funcionava, en els casos en que alguns nú-

meros o lletres no quedaven del tot ben definits o quedava encara soroll en aquella part, la lectura no era bona.

Primer s'aplica el filtre de mediana *MeidanBlur* per a acabar de netejar el soroll restant i allisar i definir més les lletres (s'han provat altres filtres també, però aquest seguia sent el que donava millors resultats). El filtre de mediana "recorre tots els elements de la senyal (en aquest cas imatge) i reemplaça cada pixel amb la mediana dels seus píxels veïns (ubicats en un quadrat de veïns al voltant del pixel avaluat"[10]. Aquest s'aplica usant un total de 5 veïns per pixel. Seguidament, es re-escala la imatge a quatre vegades més gran, per a millorar-ne la qualitat i facilitar la feina d'*OCR*.

Per a ambdós casos, s'han provat diferents combinacions. Com ara quin fer servir primer que l'altre, aplicar el filtre abans i després del re-escalat, aplicar el filtre usant diferent número de veïns per pixel, o usar diferents grossàries per al re-escalat. Finalment, la combinació descrita prèviament ha estat la més òptima.

Per últim, s'ha detectat que en els casos en que les imatges tenen parts de texts o soroll just als marges, *PyTesseract* no funciona bé i retorna text sense sentit. Per a solucionar-ho, s'afegeixen deu píxels negres a totes els costats per a assegurar que els marges queden nets i no suposen un problema.

Aplicació d'*Optical Character Recognition*

Per finalitzar l'extracció de la informació només falta aplicar *OCR* sobre les regions extretes i identificar-ne el text. Com ja s'ha explicat múltiples vegades prèviament, s'usa *PyTesseract* per a dur-ho a terme. Eina que en facilita molt la tasca, ja que 'només' s'ha hagut de preparar les imatges per a que aquesta funcioni el millor possible. L'únic detall a tenir en compte és que s'ha d'especificar quina llengua s'està tractant, i si s'extreuen només números o també text. Aquest últim va suposar algun contratemps, ja que al extreure les dates aquest funcionava força malament al principi.

2.3 Netejat de les Dades

Acabada l'extracció de la informació de les imatges, cal processar el que no és més que text en brut.

Data

El format de text de les dates correspon a *dd.mm.aaaa, hh:mm Uhr* (e.g. 31.12.2015, 23:30 Uhr), i s'han de convertir a format *datetime* per a poder treballar amb elles posteriorment. Així com la conversió en sí és fàcil, utilitzant una llibreria com pot ser *dateutil*[12], les irregularitats i soroll en el text extret suposen un problema. El format de la data extreta requereix precisió en tots els elements, sinó, o bé la data és errònia o el mètode de conversió no funciona. En els casos que els números en sí no s'han pogut llegir correctament no hi ha massa a poder fer, de manera que m'he centrat a filtrar aquelles parts de soroll que es pugui. Entre elles: assegurar que la data en forma de text és de la llargada correcte, sinó el *parser* completa la resta amb una hora i dia incorrecte. I filtrar que els signes de puntuació usats siguin els correctes, sinó el mètode de conversió no funciona correctament.

[12] *https://dateutil.readthedocs.io/en/stable/ [consulta: 11 de juny del 2018].*

Informació Gènere Oficial de Policia

El format de text dels oficials de policia encarregats del informe correspon a *Nom, Rang, Oficina* (e.g. Granada, PHK, EG Köln). L'objectiu és extreure'n el gènere, i per a fer-ho, s'aprofita que el sistema d'escriptura alemany sigui rigorós amb el gènere. Ja que el rang de l'agent ve acompanyat d'un *in* en cas de gènere femení (e.g. PHK - Home, PHKin - Dona). En primera instància, només comprovava si la segona posició del text contenia la dupla *in* o no amb distància 1. Però això comportava un problema, i és que no tenia en compte quan el text havia estat llegit incorrectament i no corresponia al agent de policia. Com a solució, es comprova que el rang contingui la lletra "P"o "K", presents a tots els rangs. El gènere es guarda en una columna extra amb un número enter, 0 per a home, i 1 per a dona.

Descripció dels fets

La descripció dels fets és extreta en un blocs de text que ja inclouen tot el necessari. Tot i això, en borro certes paraules en concret que són comuns en la majoria de texts, ja que formen part de peus de pàgina, signatura al final de la descripció, etc. A part d'això, la resta és bàsicament preprocessat de text per tal de poder analitzarlo posteriorment. S'eliminen tots els signes de puntuació, números, *stopwords*[13] i es guarda el text netejat en una columna nova. A més a més, també es crea una altra columna que conté el text netejat, però amb l'afegit que s'aplica *stemming*[14] utilitzant el *SnowballStemmer* de la llibreria *NLTK* de python. *NLTK* "és una plataforma líder per desenvolupar programes en Python que treballin amb dades de llenguatge humà. Proporciona interfícies fàcils d'usar per més de 50 corpus i recursos lèxics com el WordNet, juntament amb llibreries de processament de text per classificació, tokenització, stemming, tagging, parsing, i raonament semàntic"[13].

[13] *"Quan parlem de stop words o paraules buides ens referim a totes aquelles paraules que manquen d'un significat per sí soles. Les paraules buides solen ser articles, preposicions, conjuncions, pronoms, etc..."*[11]

[14] *"Stemming normalment fa referència a un procés heurístic aproximat que talla els finals de les paraules amb l'esperança d'aconseguir l'objectiu [reduir les paraules derivades a una base comuna] correctament la majoria de les vegades, i habitualment inclou l'eliminació d'afixos derivats."*[12]

Tipus de crim

El format dels tipus de crim consta d'una llista de frases que contenen la descripció en text del tipus d'incident i el codi associat (e.g. *Diebstahl (Par. 242 StGB)*). Es tracta d'una llista, perquè per a cada incident poden haver passat diferents delictes, tals com agressió i furt, per exemple. Com que el text és més llarg i és més susceptible a petits errors en la lectura que en dificultin un posterior agrupament per tipus, s'aprofita l'estructura que té el codi per a extreure'n el número més eficaçment. D'aquesta manera, es fa un recorregut per totes les línies del text, i es busca els tags *Par.* i *StGB* per identificar el número. Una vegada trobat, aquest s'afegeix al string que els conté tots, i es guarda. Per a poder filtrar posteriorment per text, s'ha confeccionat un diccionari a mà amb els principals tipus de delicte i el seu codi. De manera que, en cas que aquest s'hagi extret satisfactòriament, es guarden tots amb el mateix nom.

Localització

Finalment, es converteix la localització extreta en text a les coordenades de latitud i longitud. Per a fer-ho, es fa servir la llibreria *geopy*, usant el servidor *Nominativ*. Geopy "és un client per Python 2 i 3 per diferents serveis web populars de geocodificació. Geopy fa fàcil per a desenvolupadors Python localitzar coordenades d'adreces, ciutats, països, i punts de referència arreu del món usant geocodificadors de tercers i altres fonts de dades"[14]. Al extreure les coordenades de tots els elements, com que en són molts, s'intenta respectar el tràfic del servidor. De manera que s'afegeixen pauses de quatre segons entre cada petició. Tot i això, a vegades no funcionava correctament, així que es canvia al servidor *GoogleV3* temporalment. Per altra banda, el text a vegades no era llegit correctament i tenia imperfeccions, les quals causaven error al intentar retreure les geocoordenades, perquè *Nominativ* no funciona quan l'adreça no és precisa. Per a aquests casos, també es fa servir

la API de GoogleMaps, que tot i que hi hagi errors en el text de la petició, automàticament en fa la predicció. Això porta a que d'alguns texts sense cap ni peus, se'n predigui les coordenades igualment i resulti en coordenades errònies per tot el mapa terraqüi. Encara que la quantitat no és significativa, i es soluciona filtrant per geocoordenades que siguin d'Alemanya. El fet d'usar diferents servidors per a extreure les geocoordenades pot provocar imprecisions. Degut a que mateixes direccions poden tenir coordenades lleugerament diferents, però em sembla que és més profitós tenir-ne més tot i aquest petit biaix.

2.4 Dataset

El dataset resultant es guarda en format *HDF5*[15], que permet guardar grans quantitats de dades de manera eficient per a python. El fitxer se separa en dos datasets. El primer, anomenat *df*, conté la informació tal i com ha estat extreta. I el segon, anomenat *df_ clean*, conté la informació després del netejat. A més a més, també es guarden ambdós en format *csv* per a que sigui més fàcil visualitzar-ne el contingut. Per altra banda, el dataset ha estat variant en funció de les necessitats de l'anàlisi, de manera que les columnes que tenia han anat canviant fins al final.

La versió final del dataset, una vegada netejat, conté les següents columnes:

[15] *https://support.hdfgroup.org/HDF5/ [consulta: 24 de maig del 2018].*

Figura 23: Descripció de les columnes del dataset.

Columna	**Descripció**
Aufnahmezeit	Data i hora de la denúncia.
Tatzeit_am	Data i hora del incident.
Tatzeit_bis	Data i hora de la finalització del incident.
Sachverhalt	Descripció dels fets.
Sachverhalt clean	Descripció dels fets amb netejat de signes de puntuació i *stopwords*.
Sachverhalt clean stem	Mateix que *Sachverhalt clean* amb aplicació extra de *stemming* de les paraules.
Straftaten	Text brut que llista els tipus de delicte.
Straftaten_code	Netejat dels tipus de delicte, en conté el codi.
Straftaten_text	Netejat dels tipus de delicte, en conté el text.
Tatort	Adreça del incident.
Tatortlichkeit	Lloc del incident.
Tatort_lat	Latitud de les coordenades del incident.
Tatort_lng	Longitud de les coordenades del incident.
Aufnahme_ges	Text brut que conté l'oficial a càrrec de la denúncia.
Beamte-in	Gènere de l'oficial a càrrec de la denúncia.
Gegeben	*Flag* que indica si el report és buit o no.

Nota: Els noms de les columnes estan escrits en Alemany i corresponen als noms de les etiquetes dels apartats dels informes policials.

Com que al principi no se sabia exactament quins eren els límits de l'anàlisi, es va extreure tota la informació que semblava que podria ser útil. Però al final, les columnes *Tatzeit_bis* i *Tatortlichkeit* no són usades.

Percentatges d'extracció correcte de la informació

Per a tenir confiança de que els anàlisis posteriors són significants, es comprova, tal i com es mostra a la Figura 24, quin percentatge de tots els informes proporcionats la informació n'és extreta exitosament. Previ a la meva participació en el projecte, ja s'havia treballat en alguns aspectes d'aquest. Per això, ja hi havia un dataset fet per ells amb certes parts extretes a mà. Basant-se en aquest dataset, es pot saber amb seguretat que hi ha un total de 1086 reports amb informació disponible.

Figura 24: Percentatges d'informació extretes correctament. Incloent combinació de columnes.

Columna(es)	**Mostres extretes de les 1086 totals**	**Percentatge sobre total de reports**
Data de Denúncia	895	82%
Data dels Fets	853	78%
Latitud i Longitud	766	70%
Descripció dels Fets	935	86%
Tipus d'Incident	683	62%
Gènere Oficial de Policia	876	80%
Data Denúncia + Data Fets	818	75%
Lat/Long + Data Fets	698	64%
Data Denúncia + Data Fets + Lat/Long	674	62%
Data Denúncia + Data Fets + Gènere Policia	754	69%

Els apartats dels quals considerava que era més important tenir-ne un bon número de mostres, eren les Dates de Denúncia i les dels Fets. Perquè la part més important de l'anàlisi es centraria en observar si hi ha patrons en les descripcions dels fets en funció de la diferència entre ambdues dates. En primera instància, se'n

tenia només un 23% del total. No n'era prou, així que es va millorar la conversió de text a data, però tot i això se seguia tenint només un 32% (66% i 39% respectivament per separat). De la resta de columnes, els percentatges d'extracció no eren gens dolents en comparació, de manera que vaig decidir modificar les dates que no s'extreien bé manualment. Imprimia una llista amb totes les dates incompletes, i les afegia. Encara que no el 100% d'aquestes, sinó que només aquelles de reports estàndard i que s'hagués extret en certa mesura part d'alguna de les dues dates. Aquest és un pas que, sinó feia jo en aquest moment, haurien fet els coordinadors del projecte en el futur, així que em va semblar millor fer-ho i donar-li més substància al treball.

Quantitat d'informes estàndard i no estàndard

Una vegada finalitzat el dataset, és interessant saber exactament quin número de reports estàndard i no estàndard hi havia. Per a ratificar la decisió de només haver-se centrat en els primers. Amb aquest fi, s'ha reaprofitat la CNN construïda prèviament per a fer la classificació. S'escullen pàgines d'un i altre tipus per a l'entrenament, es recorren totes les carpetes agafant la primera pàgina del report i se la classifica.

D'entrada, es comptaven un total de 838 informes estàndard i 238 de no estàndard. Però si mirem només la quantitat de Dates de Denúncia (895) veiem que no és possible, ja que aquestes només es poden extreure dels estàndard i segons la CNN n'hi hauria més. S'ha intentat fer retocs a la xarxa per a aquest cas també, però novament sense resultats.

Finalment, s'aprofita el fet que dels reports no estàndard no es pot extreure tota la informació per a fer el recompte. Amb aquest mètode, es miren tots els informes donats no buits, i es compten tots els que tenen la majoria de columnes que requereixen processament (Dates, Localització, Gènere del/a Policia, etc.) nul·les.

En total, s'aproxima que hi ha 135 reports no estàndard o dels que la qualitat del escanejat és tan dolenta que no se'n pot extreure la informació. El que deixa amb 951 informes estàndard.

3 Anàlisis de Dades

L'objectiu principal que tenien els coordinadors del projecte era el d'extreure les dades dels informes escanejats, sobretot les parts de la descripció dels fets i dates. Tanmateix, evidentment l'objectiu final és analitzar aquestes dades, especialment buscant canvis en els reports en funció del temps que triga la víctima a denunciar, o trobar paraules clau al llarg dels temps. En els següents apartats es descriuen les eines desenvolupades per a analitzar les dades. Cal remarcar que aquestes no estan pensades per a ser les finals, sinó com a primera aproximació al problema per als membres del projecte, els quals saben millor exactament en què volen centrar l'anàlisi per a treure'n conclusions. Tot el codi creat per a l'anàlisi es troba a la carpeta *Analyse* del codi entregat. Per altra banda, totes les figures, mapes i estadístiques resultants es troben a la carpeta *Results*.

3.1 Geolocalitzcions

S'ha creat codi que permet visualitzar en mapes de calor les geolocalitzacions dels incidents. L'objectiu és observar quines zones concretes de la ciutat han estat més conflictives, i alhora fer-ho amb un format visual que en faciliti l'anàlisi. A més a més, és oportú que, amb fàcils modificacions, es pugui filtrar el dataset per segons quines condicions. I així poder considerar les geolocalitzacions en funció del tipus de crim, dates, etc.

La llibreria usada és *Folium*[16], la qual permet integrar dades manipulades en python i visualitzacions amb la llibreria *Leaflet*[17]. D'entrada, es va treure la idea i la base del codi per fer servir aquesta llibreria del blog *How to: Folium for maps, Heatmaps & time data*[18]. Després, s'han fet les modificacions i extensions neces-

[16] *https://python-visualization.github.io/folium/ [consulta: 19 de juny del 2018].*
[17] *https://leafletjs.com/ [consulta: 19 de juny del 2018].*
[18] *https://www.kaggle.com/daveianhickey/how-to-folium-for-maps-heatmaps-time-data/ [consulta: 19 de juny del 2018]*

sàries per adaptar-ho als requeriments volguts. Es fan servir dos tipus de mapes diferents: *HeatMap*, que permet crear un mapa de calor, i *HeatMapWithTime*, que, donades les dades agrupades per valors, permet fer una visualització del mapa de calor en funció de les dates. El mapa base emprat és sempre *OpenStreetMap*, que és un mapa del món obert i gratuït construït per la comunitat. [19]

El codi realitzat genera arxius *html* que contenen els mapes. Cal tenir en compte que *folium* els carrega d'internet cada vegada. Per apreciar la interactivitat dels mapes, es poden obrir i fer servir. Es troben a la carpeta *Maps* dins de *Results* del codi entregat. Es poden crear tres tipus de mapes diferents:

1. Única plana amb totes les localitzacions. Per a aquestes visualitzacions, només s'han d'agrupar les geolocalitzacions de tot el dataset donat i crear el mapa amb elles. En la Figura 25 es mostra un exemple.

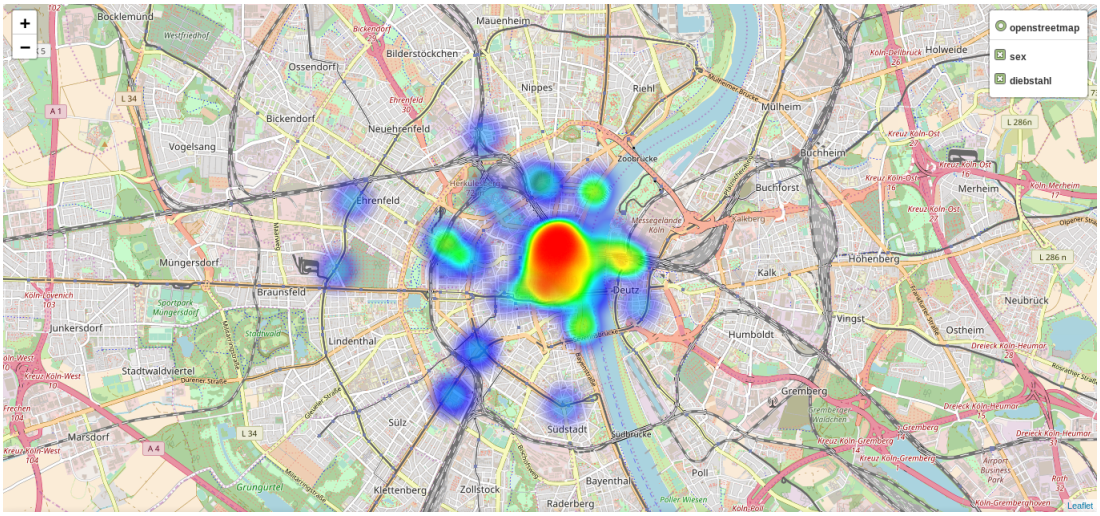
Figura 25: HeatMap amb totes les localitzacions.

2. Única plana amb diferents capes, de manera que es puguin sobreposar les localitzacions de diferents filtres. A la caixa situada a la cantonada superior dreta, es seleccionen les capes que es volen visualitzar. Per a aquestes visualitzacions, per a cada capa s'ha de filtrar el dataset i preparar les dades tal i com es fa en el primer tipus. Inicialment el codi només permet filtrar per

[19] *https://www.openstreetmap.org [consulta: 19 de juny del 2018].*

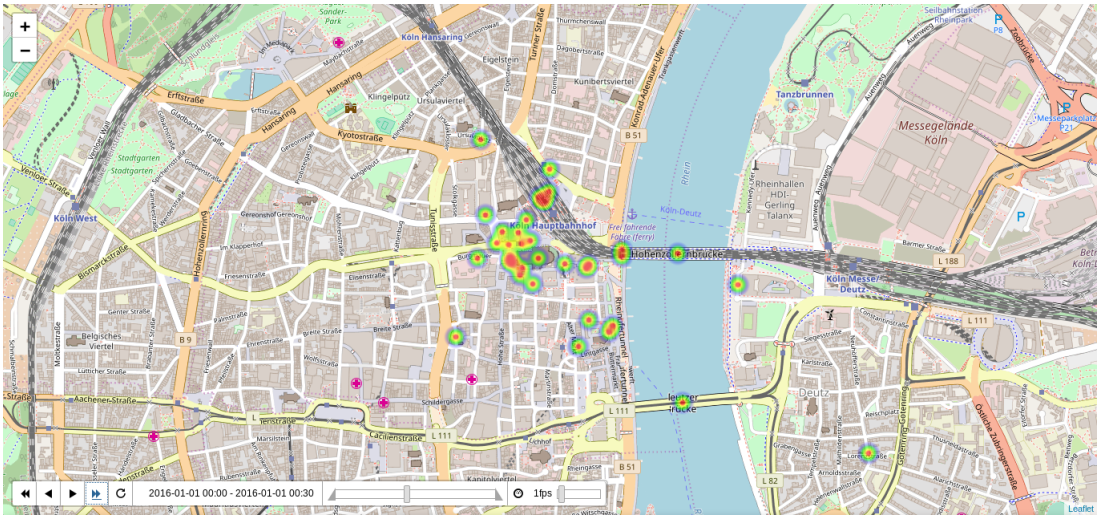
tipus d'incident, però és fàcilment modificable per a filtrar per a qualsevol altre apartat del dataset. En la Figura 26 es mostra un exemple.

Figura 26: HeatMap amb diferents capes.



3. Múltiples planes agrupades per valor d'una columna concreta del dataset, especialment pensat per a visualitzar les localitzacions per dates. S'usa la barra inferior esquerra per a moure's entre planes. Per a aquestes visualitzacions, s'agrupen les dades en funció de la columna desitjada. De manera que per a cada valor diferent de la columna, es fan servir les geolocalitzacions corresponents per a crear la plana. En la Figura 27 es mostra un exemple.

Figura 27: HeatMap amb tots els incidents separats per l'hora de l'esdeveniment.



Emprant aquest codi, es poden crear tants mapes com es vulgui filtrant el dataset

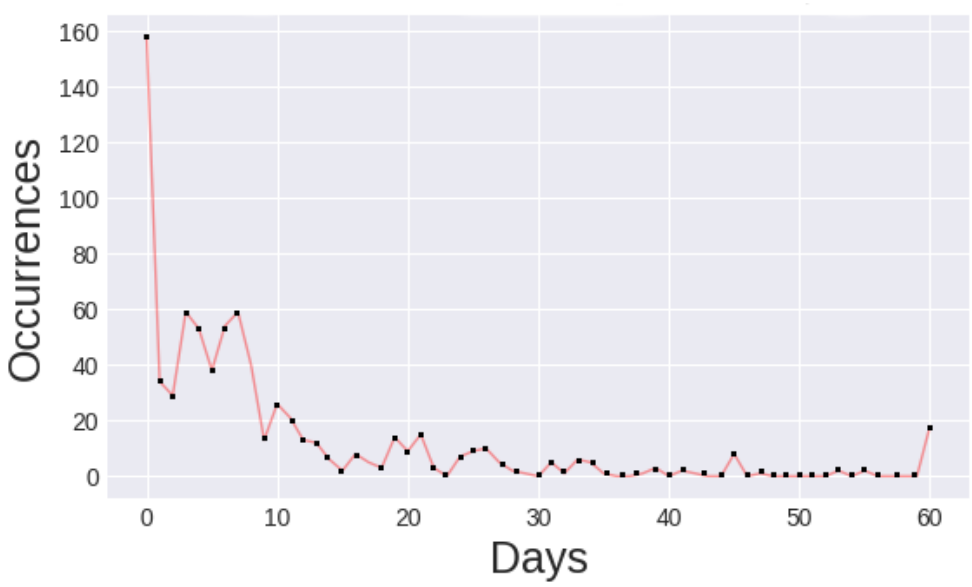
prèviament a plaer. Només cal passar el dataset filtrat i/o les columnes per a agrupar posteriorment. D'entrada i com a exemple, s'han creat els següents mapes:

- Un mapa del tipus 1 amb tots els incidents junts, anomenat *all_ locations.html*.
- Un mapa del tipus 3 agrupant pels dies que ha trigat la víctima a denunciar, anomenat *aufnahme_ days.html*.
- Un mapa del tipus 3 agrupant per hora de l'incident tots els robatoris, anomenat *diebstahl_ time.html*.
- Un mapa del tipus 2 que conté dues capes, una per als incidents amb trets sexuals i una per a robatoris, anomenat *filtered_ incidents.html*.
- Un mapa del tipus 3 agrupant per hora de l'incident tots els que tenen trets sexuals, anomenat *sexuelle_ grund_ time.html*.
- Un mapa del tipus 3 agrupant per hora tots els incidents, anomenat *tatzeit_ am.html*.

3.2 Temps que triguen les víctimes a denunciar

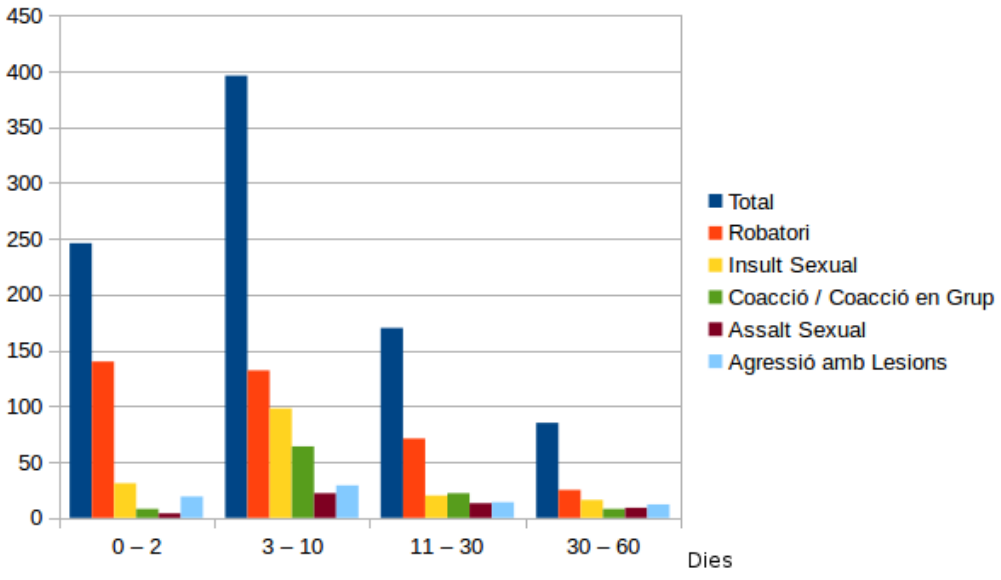
Un dels tòpics principals del projecte és el de revisar si hi ha patrons respecte el temps que triguen les víctimes a denunciar. Analitzar si hi ha canvis significatius en les descripcions, paraules clau, expressions, to de la víctima o si algun tipus de crims es triguen més o menys a denunciar. Tots els relatius a text seran tractats en els següents apartats. Per altra banda, a continuació, en la Figura 28, es mostren la quantitat de denúncies que hi ha per cada dia posterior al incident. Cal tenir en compte però, que això no indica el número d'informes individuals per dies, sinó el número de tots els tipus d'incidents que han passat per dies. Ja que es compten diferents crims d'un mateix informe per separat. Es pot apreciar com hi ha una gran quantitat de denúncies les primeres 24 hores i, a partir d'aquí, hi ha una baixada radical el mateix dia 1, seguit per una baixada gradual al llarg dels altres dies. Finalment, a partir de més de 30 dies n'hi ha de molt poques al dia.

Figura 28: Número d'incidents per dies transcorreguts fins la denúcia.



Tal i com s'ha comentat, resulta interessant analitzar si hi ha algun tipus d'incident que es triga més a denunciar. A la gràfica de la Figura 29 es visualitzen els tipus d'incidents més significatius i el número que n'hi ha entre transcorreguts de 0 a 2 dies, 3 a 10 dies, 11 a 30 dies i més de 30 dies. Comparant-los alhora amb el total d'incidents en els mateixos rangs per a comparar-ne proporcions.

Figura 29: Gràfic que agrupa per blocs de dies els tipus d'incidents.



A la Figura 28 es veu com quan hi ha més denúncies és en les primeres 24 hores. Tot i això, a la Figura 29 s'aprecia com si s'agrupen dels 3 als 10 dies posteriors a l'incident en total n'hi ha més. Les discussions d'ambdues figures es duen a terme al capítol següent, *Discussió dels Resultats*.

3.3 Text

En aquest apartat, s'estudia tot el que té a veure amb el text. Buscar patrons, termes freqüents, paraules clau o analitzar quin tipus de paraules són usades al llarg del temps, entre d'altres. Inicialment es va voler aplicar Latent Dirichlet Allocation (LDA) model que busca diferents tòpics en un conjunt de texts basant-se en la freqüència de les paraules, com a primera aproximació per a trobar patrons. Es va seguir el tutorial de Jordan Barber[15] que usa la llibreria *Gensim*[20]. Aplicar-ho permetia confirmar que hi havia dos tipus de texts predominants: la descripció dels fets escrita pels oficials, i la mateixa essent una transcripció literal de les víctimes. Al aplicar LDA, en el primer grup predominaven els verbs en tercera persona i les paraules GES i BES. Aquestes dues corresponen a *Geschädigt* (víctima) i *Beschädigungen* (danys), usades al fer la descripció per un oficial. A més a més, aquestes apareixien entre les paraules més comuns, el qual significa que n'hi ha molts exemplars. Per altra banda, en el segon grup apareixia majoritàriament la primera persona i contenia força més adjectius que el primer, del que es pot induir la separació comentada. Tot i això, aviat es va deixar de fer servir perquè no semblava que se li pogués treure més suc.

A continuació, s'han fet gràfics contenint els 100 unigrames, bigrames i trigrames més comuns, també els 100 noms més comuns i els 100 noms més comuns a partir dels 10 dies. Per a fer-ho, s'ha usat la llibreria NLTK, que és la principal llibreria Python per a processament de text. Desgraciadament, en el seu moment no es va pensar en que s'haurien d'incloure en la memòria i resulten massa grans per a apreciar-los apropiadament en la memòria. Tanmateix, aquests es troben a la

[20] *https://radimrehurek.com/gensim/index.html [consulta: 22 de juny del 2018].*

carpeta *Text* dins de *Results* del codi entregat, en cas que es vulguin visualitzar en foto. Per altra banda, traduir el contingut dels gràfics al català no sembla oportú per la pèrdua de matisos de la llengua. A continuació, es mostren els resultats en les Figures 30-33.

Figura 30: Gràfics amb els 100 noms més comuns.

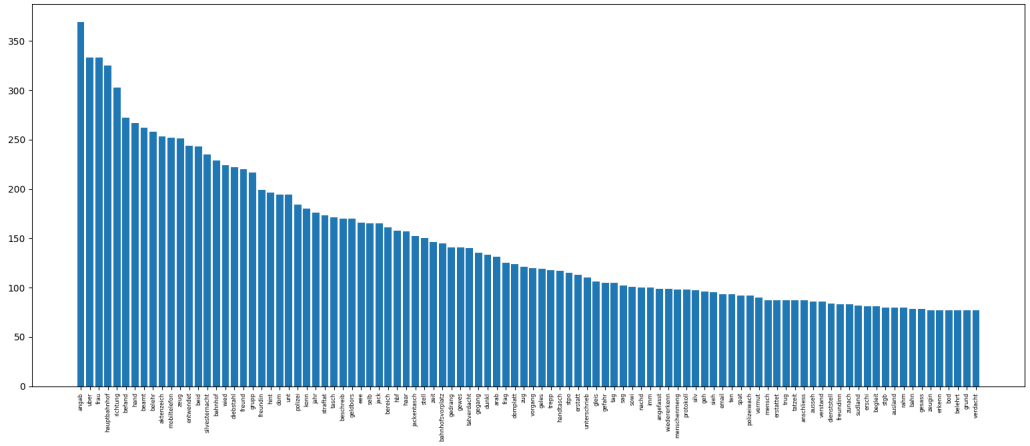


Figura 31: Gràfics amb els 100 noms més comuns a partir dels 10 de la denúncia.

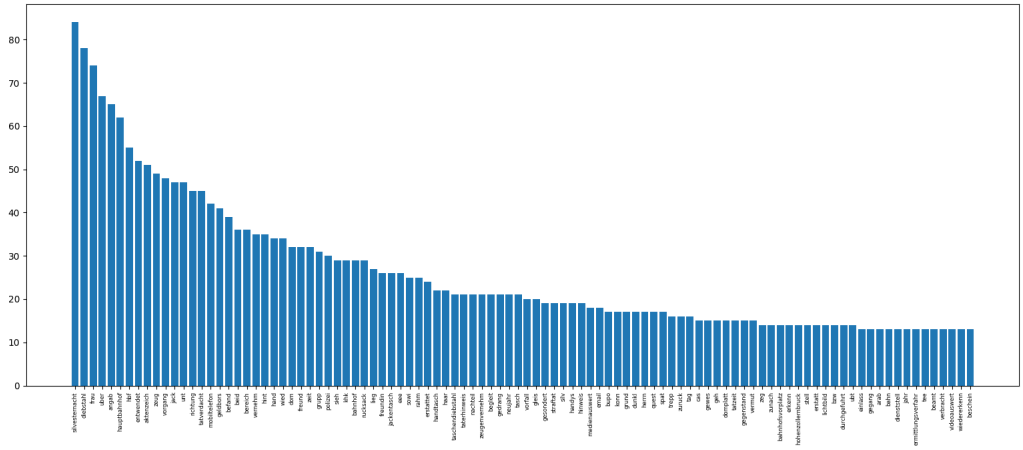


Figura 32: Gràfics amb els 100 bigrames més comuns.

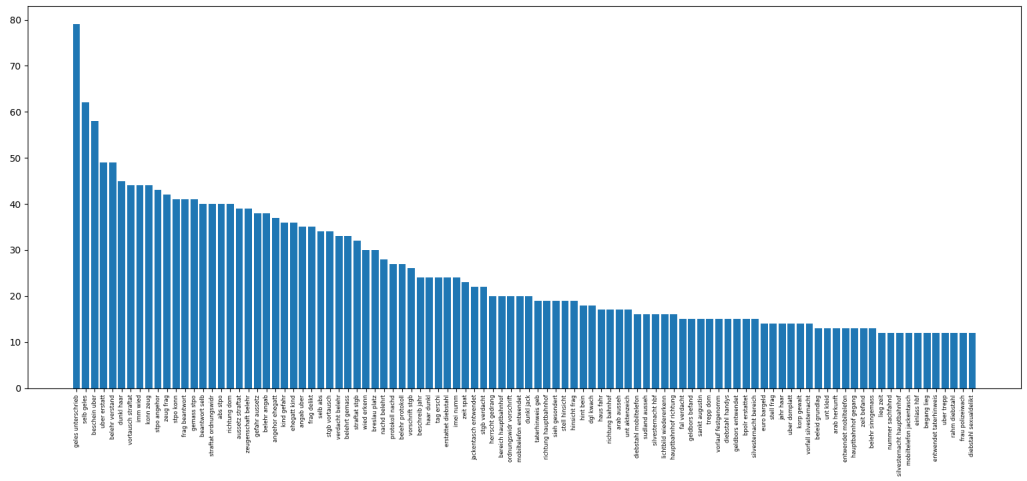
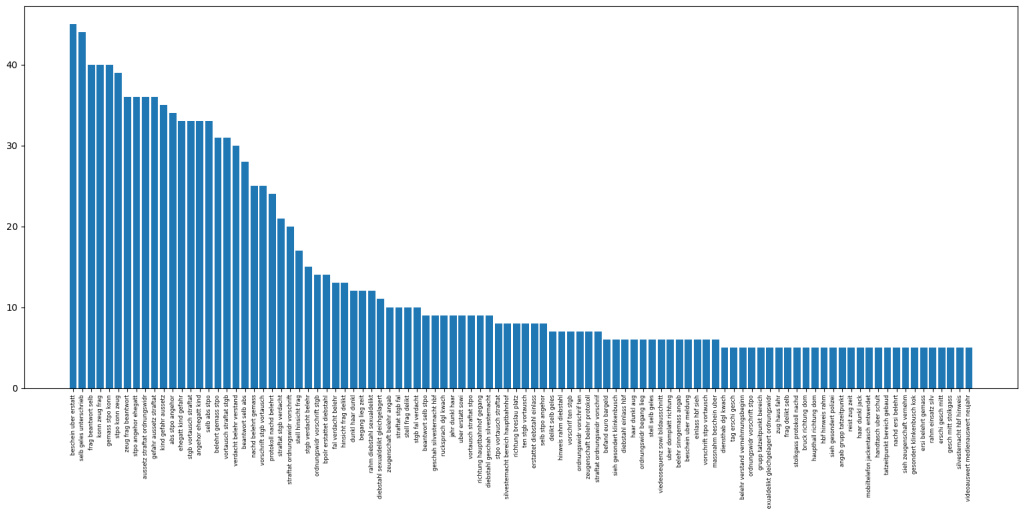


Figura 33: Gràfics amb els 100 trigrames més comuns.



El dataset pot ser filtrat per una llista de paraules qualsevol, de manera que només quedin els informes que les incloguin. Per a usar-ho, s'ha demanat una llista preliminar de paraules al departament de psicologia:

1. Gedränge - Aglomeració
2. Vergewaltigung - Violació
3. Sexuell belästigt - Assetjament sexual
4. Umzingelt - Envoltat

5. Kontrolle - Control
6. Antanzen - Aparèixer
7. Angst - Por
8. Gruppe - Grup
9. nordafrikanisch - Nord-africà
10. gewalttätig - Violent
11. Aggressiv - Agressiu
12. Belästigung - Assetjament
13. Berührt - Tocat
14. Masse - Apinyament
15. Fremd - Estrany

Aquestes són considerades paraules clau interessants per a observar com de freqüents són, sobretot en relació al que triguen les víctimes a denunciar. Com que posar moltes paraules alhora al mateix gràfic en dificulta la comprenció, aquestes es separen en tres grups de cinc paraules cada un. Principalment, hi ha tres tipus de gràfics: un primer que mostra les freqüències per tots els dies, un segon que les mostra només a partir dels 10 dies, i un tercer que les mostra en relació al total d'incidents de cada dia. A més a més, amb el mètode desenvolupat, resulta fàcil crear nous gràfics amb altres paraules que es vulgui a plaer. A continuació es presenten els gràfics usant les paraules anteriorment presentades:

Figura 34: Freqüència de les paraules 1-5 respecte els dies que es triguen a denunciar.

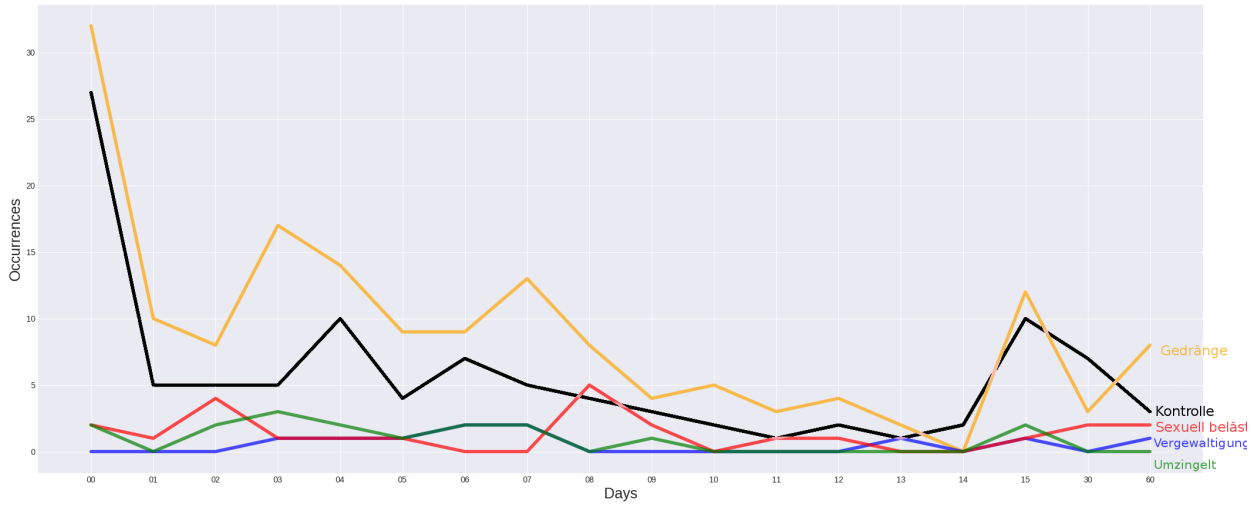


Figura 35: Freqüència de les paraules 1-5 a partir dels 10 dies de la denúncia.

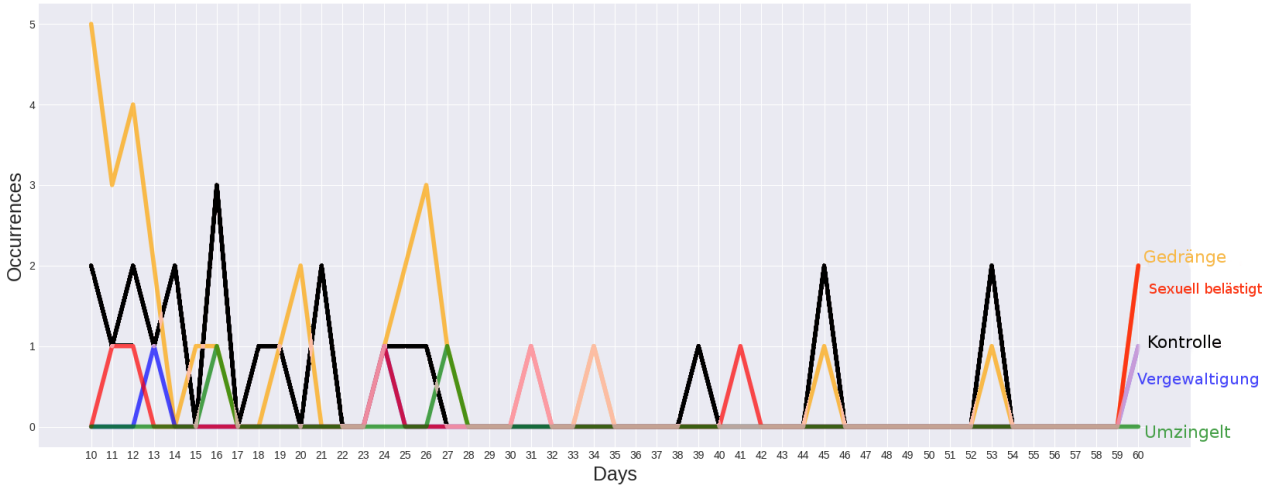


Figura 36: Freqüència de les paraules 6-10 respecte els dies que es triguen a denunciar.

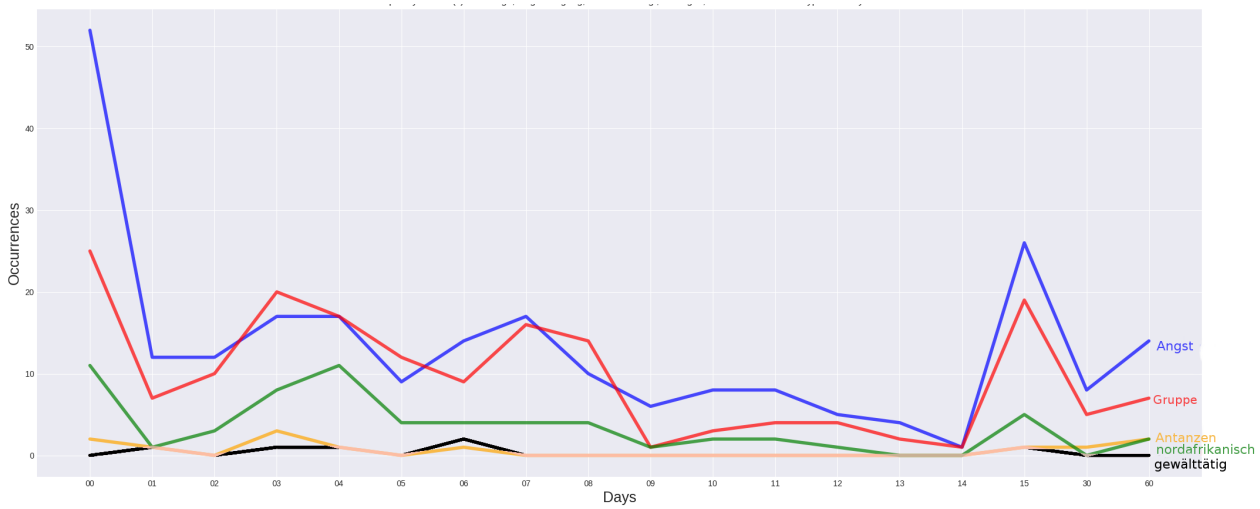


Figura 37: Freqüència de les paraules 6-10 a partir dels 10 dies de la denúncia.

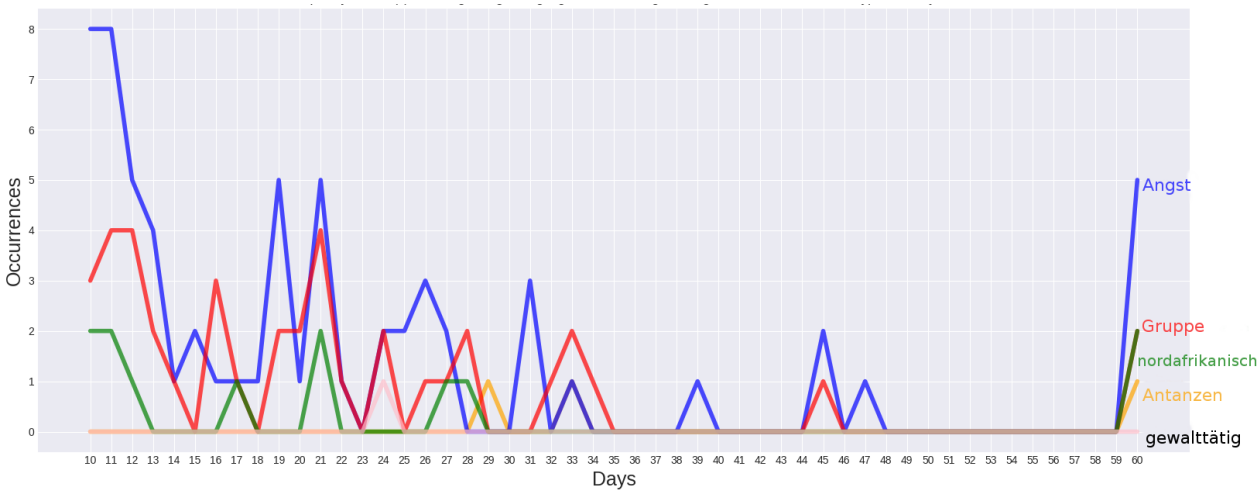


Figura 38: Freqüència de les paraules 11-15 respecte els dies que es triguen a denunciar.

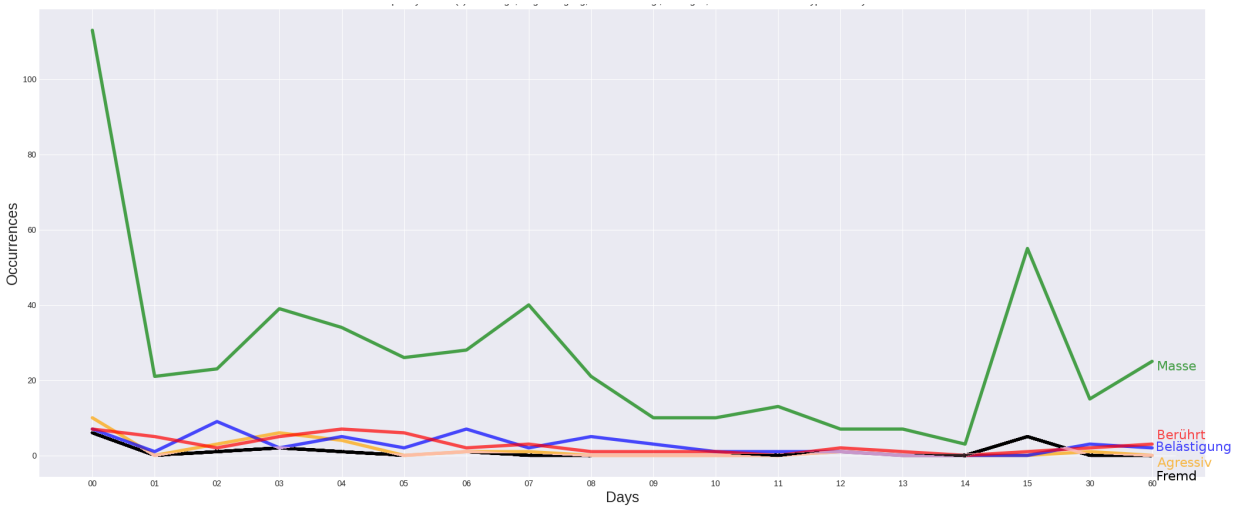
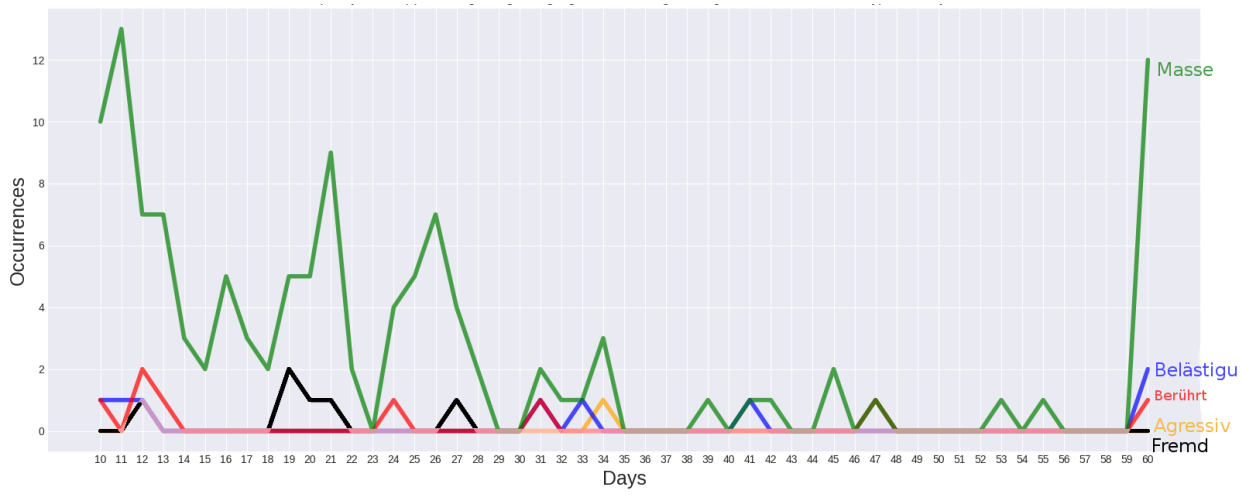
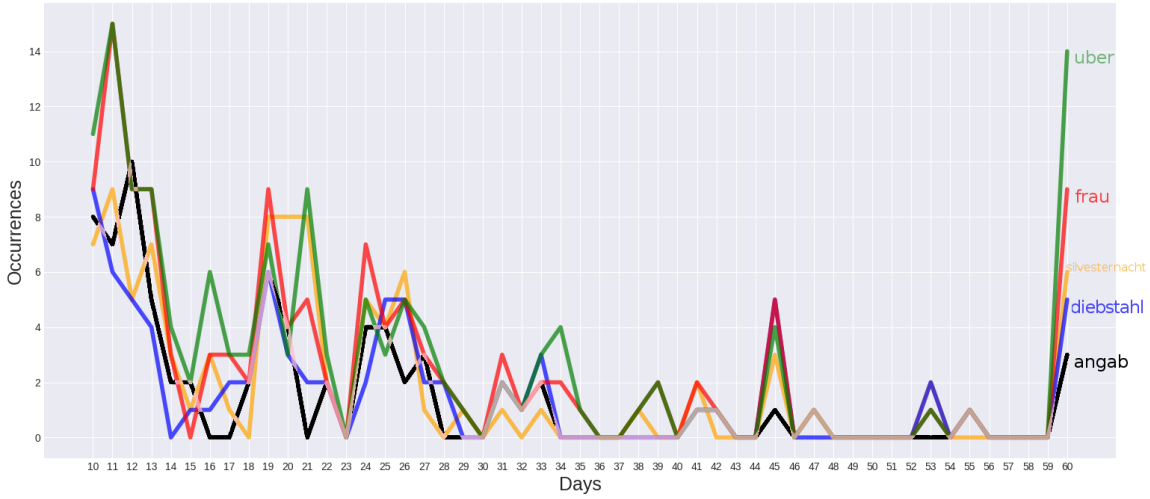


Figura 39: Freqüència de les paraules 11-15 a partir dels 10 dies de la denúncia.



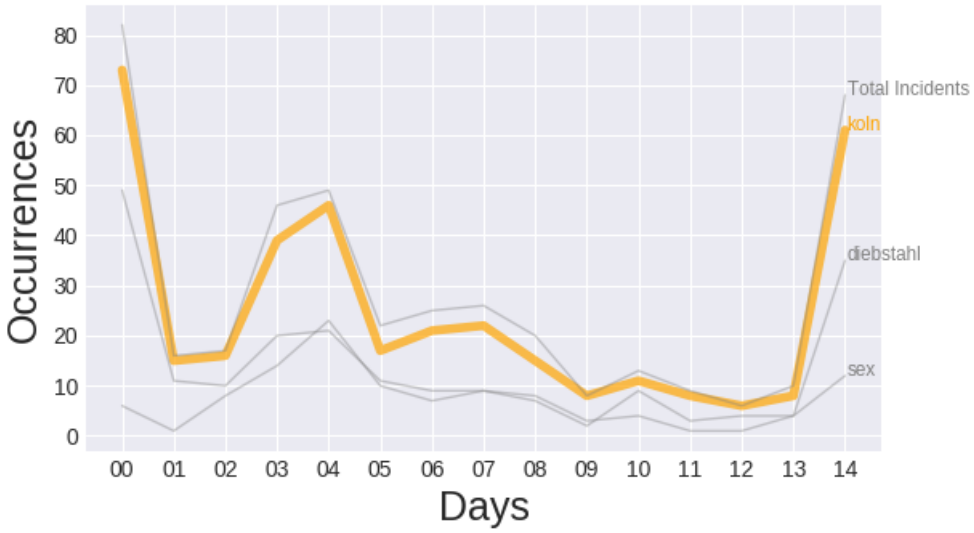
A més a més, també es va proposar observar les freqüències de les 100 paraules més comuns a partir dels 10 dies transcorreguts entre l'incident i la denúncia. A continuació, a la Figura 40, se'n mostra un exemple dels 20 gràfics hi ha. Altrament, n'hi ha 20 més amb les mateixes condicions però afegint el total d'incidents com a comparativa.

Figura 40: Freqüència de les paraules 'uber' (sobre), 'Frau' (dona), 'Silversternacht' (Nit de Cap d'Any), 'Diebstahl' (Robatori) i 'Angab' (dada) a partir dels 10 dies.



Per altra banda, també es dóna la possibilitat de crear gràfics que continguin les freqüències de les paraules en relació amb les freqüències de tipus d'incidents concrets. La Figura 41, tot i no aportar informació útil com a tal, en mostra un exemple:

Figura 41: Freqüència de la paraula 'koln' en relació als robatoris (diebstahl) i incidents d'índole sexual (sex).



3.4 Sentiment Analysis

Sentiment Analysis, també conegut com a *Opinion Mining*, és "el procés de determinar si un escrit és positiu, negatiu o neutral"[16]. És un camp força complex i que encara té recorregut per a ser estudiat, especialment per a llengües diferents a l'anglès. Com que els membres del projecte estaven interessats també en buscar canvis en la polaritat de les víctimes en funció del temps que trigaven a denunciar, es va decidir usar aquesta disciplina per a fer-ne l'anàlisi.

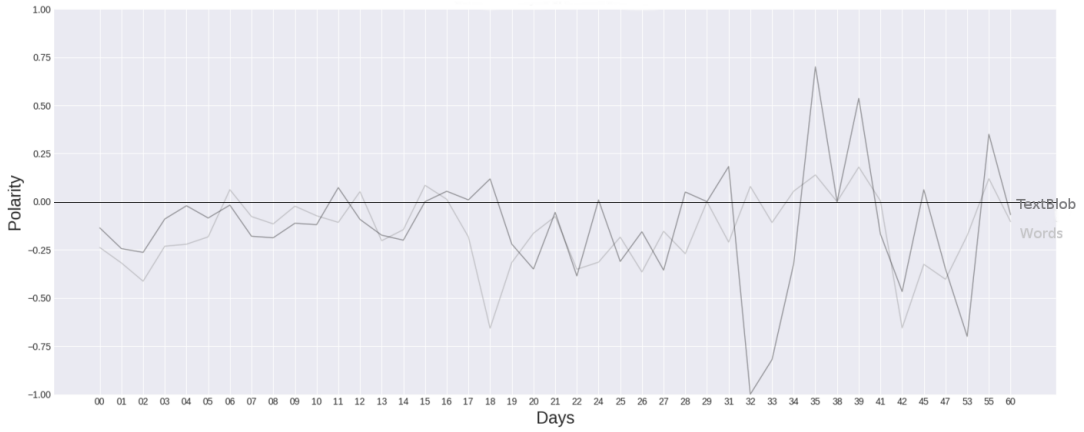
La naturalesa de les dades limita les opcions per a poder estudiar la polaritat de les descripcions. Al ser texts en alemany, les eines de software lliure no en línia escassegen (e.g. la llibreria NLTK té suport per a *sentiment analysis* només en anglès). A més a més, els estudis al respecte que no impliquin un pas d'entrenament previ tampoc són comuns. Com que les descripcions extretes no són etiquetades, no es pot entrenar un classificador, com ara una xarxa neuronal, que aprengui d'una part de les mostres. Per altra banda, entrenar usant altres mostres de text com, per exemple, ressenyes de pel·lícules, no sembla adequat per al problema.

Finalment, em vaig decantar per a emprar dos mètodes diferents. En primera instància es va provar un mètode més naïf, el qual consistia en usar una llista preconcebuda de paraules etiquetades amb un valor entre -1 i 1. Aquesta primera, que consisteix en dos fitxer diferents, un amb paraules negatives i un altre de positives, es va extreure del dataset *German Sentiment Analysis Toolkit*[17]. Inicialment, es sumaven les polaritats de totes les paraules i es dividia pel número de texts per extreure el valor donat per a un grup. El problema donat és que aquesta aproximació pot ésser massa simple, ja que no contempla la negació. Es va intentar implementar un algoritme per detectar-la en frases en alemany, però era un concepte complex per a mi i no hi va haver èxit. Posteriorment, es va trobar una llista més complexa que no conté només paraules, sinó combinacions de ngrames també. Dels quals alguns inclouen la negació o diferents formes d'aquesta. El fitxer[18] és proporcionat pel grup de reserca *Analytical Information System* del *Institute of Information Systems*

at the University Hof[21]. En darrer lloc, després usar i aplicar els mètodes mencionats prèviament, s'ha trobat una llibreria de processament de text per a python que dóna suport de *sentiment analysis* per a l'Alemany. La llibreria en qüestió és *TextBlob*[22], la qual dóna resultats més sòlids als proposats anteriorment ja que és un mètode més complet i complex.

A l'hora de comprovar la polaritat de les descripcions dels informes en funció dels dies que triguen les víctimes a denunciar, la llista llarga donava valors en un rang força diferent al dels altres dos. De manera que, al representar-ho en un gràfic provocava confusió per entendre'l i no s'inclou. Tot i això, sí que serveix com a indicador al comparar els tres mètodes posteriorment. En un inici, es representaven tots els dies individualment fins als 60, però com s'aprecia en el gràfic de la Figura 42, a partir dels 15 dies aquest té pics molt pronunciats. Això és degut a que hi ha menys mostres de reports per cada dia i els resultats són més extrems.

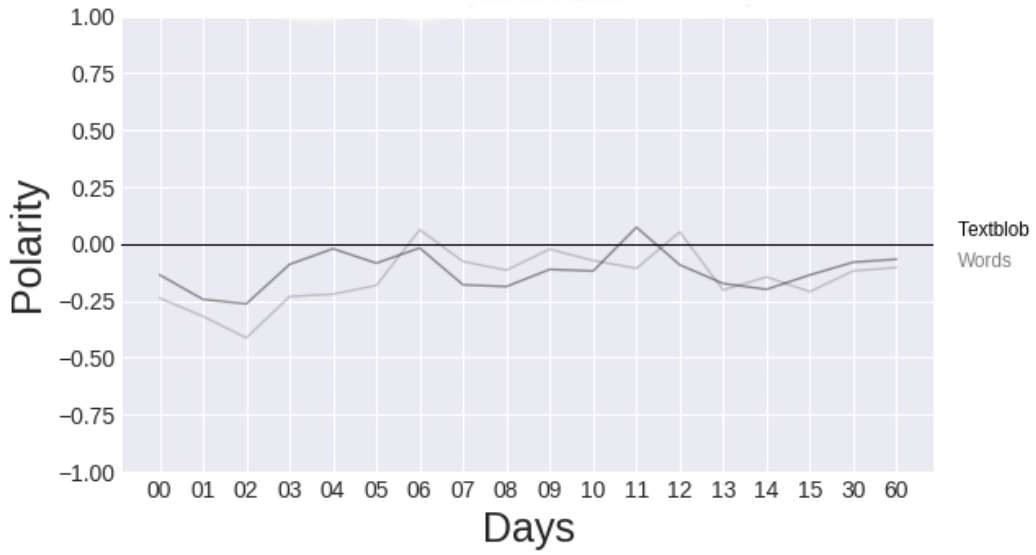
Figura 42: *Sentiment analysis* dels incidents per dies que triga la víctima a denunciar.



De manera que s'agrupen els dies del 15 al 30, del 30 al 60 i més de 60 (Figura 43. D'aquesta manera es veu com hi ha certa continuïtat i sentit.

[21] *https://www.iisys.de/en/research/research-groups/analytical-information-systems.html*
[22] *https://textblob.readthedocs.io/en/dev/ [consulta: 21 de juliol del 2018].*

Figura 43: *Sentiment analysis* dels incidents per dies que triga la víctima a denunciar.



També és important remarcar que aquests corresponen al *Sentiment Analysis* de tots els tipus d'incidents barrejats. Sense separar per robatoris, agressions, etc., que intuïtivament haurien de donar resultats diferents entre ells. Tot i que seria interessant, en el seu moment no es va considerar analitzar-ho, i com que al escriure la memòria ja no tenia accés al dataset, resulta impossible fer-ho a posteriori. Les conclusions dels resultats obtinguts es troben al capítol següent, *Discussió dels Resultats*.

3.5 Influència del gènere del/a policia

Per acabar, s'ha analitzat si hi té alguna influència el gènere del/a policia encarregat/da a l'hora de tractar amb les víctimes.

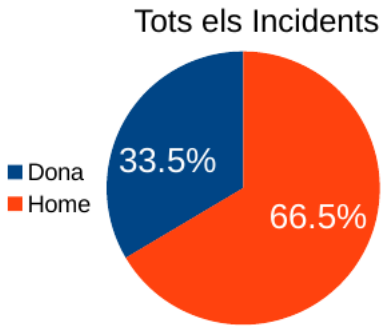
Primer de tot, es comprova els percentatges de tipus de crims tractats per policies

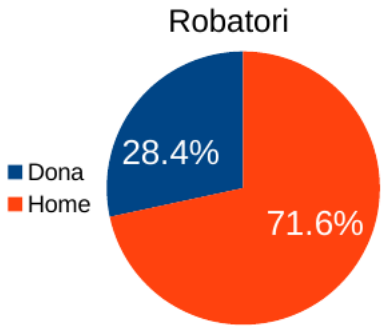
homes o dones, tal i com es mostra a la següent taula (Figura 44). La idea és observar si hi ha algun tipus de crim que normalment és tractat més per algun d'ambdós gèneres en concret. Es donen tots aquells crims que tenen suficient mostres per donar resultats concloents, la resta s'agrupa en l'apartat *altres*:

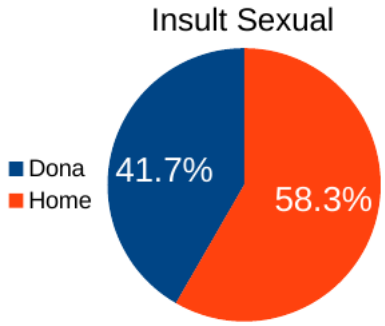
Figura 44: Total d'informes treballats separats pel gènere del/a policia.

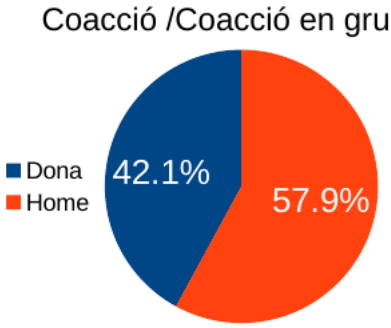
Tipus d'incident	**Dona**	**Home**	**Total**
Tots	332	658	990
Robatori	113	285	398
Insult Sexual	75	105	180
Coacció / Coacció en grup	45	62	107
Assalt Sexual	21	37	58
Agressió amb Lesions	26	59	85
Altres	52	110	162

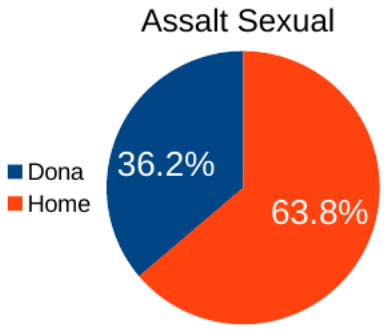
A continuació, es mostren els resultats de la taula anterior (Figura 44) amb gràfics que representen els percentatges respecte del total d'incidents de cada tipus de crim:

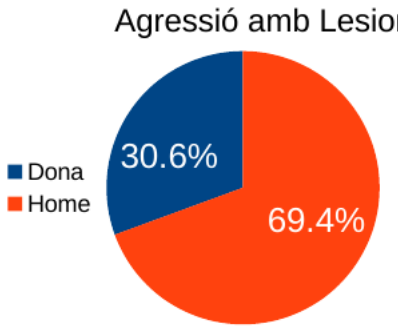

Tots els Incidents
Dona
Home
33.5%
66.5%

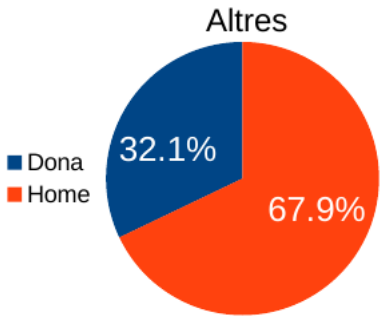

Robatori
Dona
Home
28.4%
71.6%


Insult Sexual
Dona
Home
41.7%
58.3%


Coacció /Coacció en grup
Dona
Home
42.1%
57.9%


Assalt Sexual
Dona
Home
36.2%
63.8%

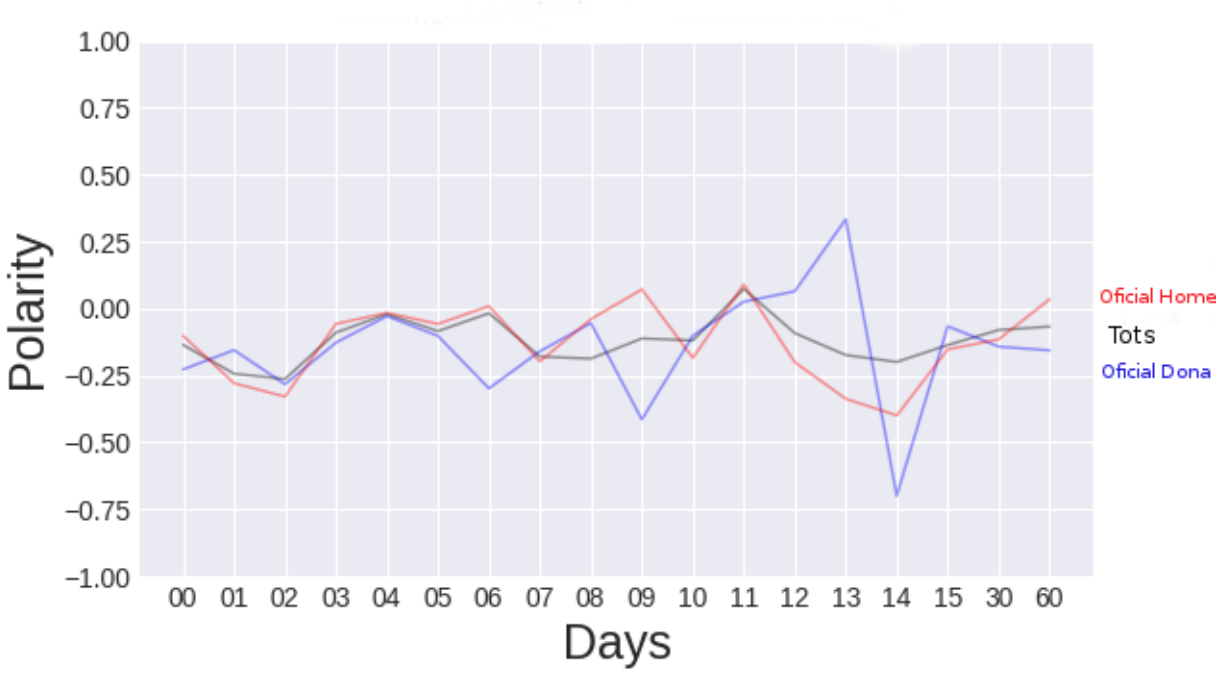

Agressió amb Lesions
Dona
Home
30.6%
69.4%


Altres
Dona
Home
32.1%
67.9%

Com a primera aproximació per veure si hi ha diferències en el text en funció de si l'agent és de sexe masculí o femení, es van extreure els unigrames, bigrames i trigrames més comuns en ambdós sexes. En els fitxers *ngrams_female.txt* i *ngrams_ male.txt* de la carpeta *LDA* inclosa en l'apartat *Text* es guarden aquests. No he trobat oportú incloure'n mostres gràfiques dels mateixos, perquè des d'un inici es va considerar que no hi havia diferència palpable, o almenys no emprant aquesta via d'anàlisi. La majoria de ngrames coincideixen i els més comuns formen part simplement de descripcions o de patrons comuns a l'hora d'escriure dels oficials.

Finalment, s'apliquen les eines de *sentiment analysis* explicades a l'apartat anterior per a tenir una altra aproximació a l'assumpte, i que ens hauria de donar una visió més sòlida. Únicament usant *TextBlob*, la Figura 45 en representa el resultat.

Figura 45: *Sentiment Analysis* dels informes en funció dels dies que triga la víctima a denunciar separats per gèneres del/a policia



4 Discussió dels Resultats

Una vegada creades i aplicades les eines d'anàlisi, cal interpretar-ne els resultats obtinguts. Tanmateix, com es veu a la taula de la Figura 24, no s'usa el 100% del dataset en totes les parts de l'anàlisi. De manera, que pot ser que la fiabilitat no sigui suficient per a que tots els resultats siguin concloents. Els membres del projecte en són conscients, i en cas necessari s'han deixat les eines necessàries per a que puguin omplir el dataset manualment i seguir treballant.

Geolocalitzacions

Si s'analitza el mapa amb tots els incidents alhora (Figura 46), s'observa com les parts més conflictives han estat al centre de la ciutat: la plaça de la catedral i carrers annexos, l'estació central, el pont i la zona de restaurants al costat del riu. Observant el mapa que conté també els afores de la ciutat (Figura 47), s'aprecia com hi ha patrons que marquen línies rectes des del centre de la ciutat, que pot significar que els agressors atacaven les víctimes ja anant cap a les festes, o tornant d'elles. Quan, a més a més, es miren els mapes en funció de les hores dels incidents, en concret entre la 1:30 i les 8:00 del matí del dia 1 (Figura 48), es veu com la majoria dels incidents a les afores es duen a terme a les acaballes de la nit, sobretot en comparació amb les altres hores, en les quals la majoria d'incidents van passar al centre. Tanmateix, dóna lloc a concloure que molts dels incidents als afores es donaren quan els agressors i les víctimes probablement estaven tornant cap a casa.

Figura 46: Geolocalització de tots els incidents al centre.

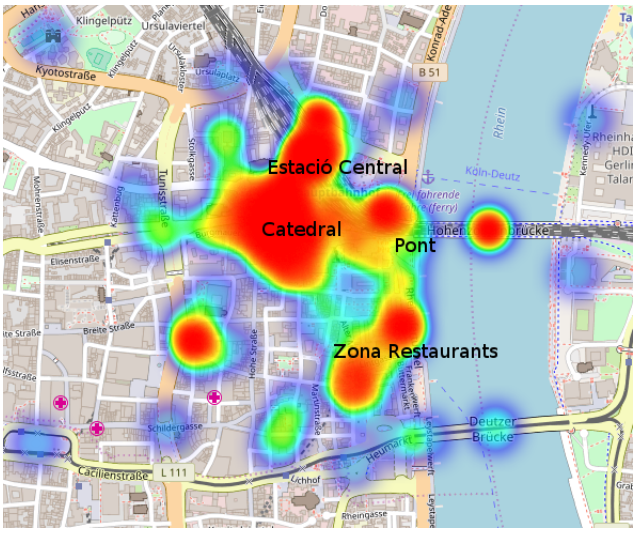


Figura 47: Geolocalització de tots els incidents a les afores en totes les franges horàries.

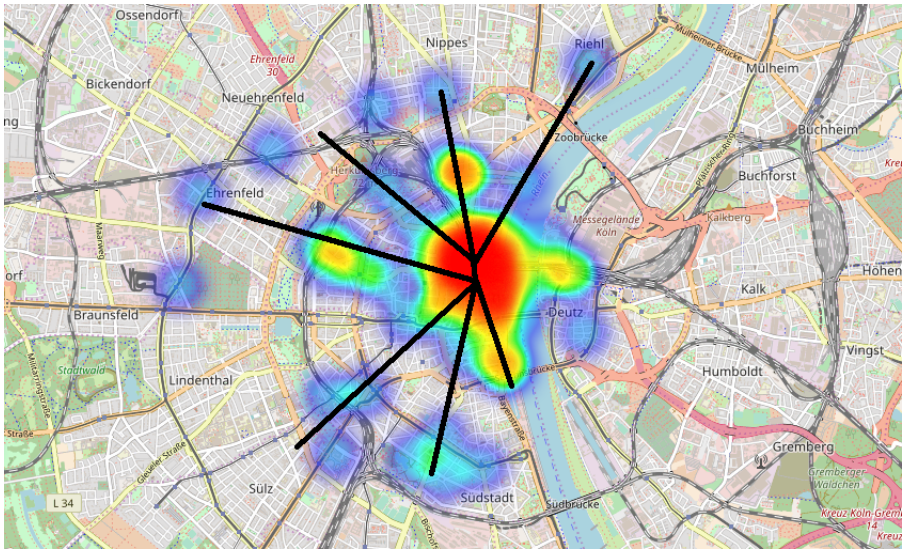
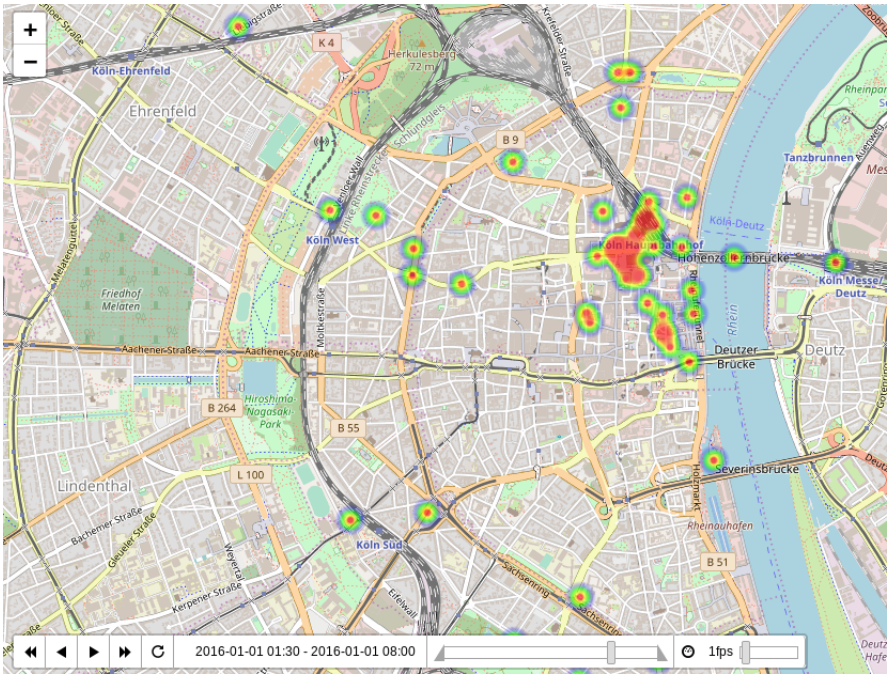


Figura 48: Geolocalització de tots els incidents entre la 1:30 i les 8:00 del 01/01/2016.



Per altra banda, si comparem els mapes amb els robatoris (Figura 49) i incidents relacionats amb agressions de caràcter sexuals (Figura 50) per separat, no sembla que hi hagi cap patró significatiu que els diferenciï. Tot i que hi ha més robatoris que agressions sexuals, les zones de la ciutat que ocupen són les mateixes, amb la simple diferència que d'un n'hi ha més que de l'altre. Tanmateix, aprofitant aquest tipus de mapes, es podrien comparar tot tipus d'incidents i altres característiques d'aquests, fàcilment superposant uns i altres, com als exemples donats.

Figura 49: Geolocalització de tots els robatoris.

Figura 50: Geolocalització de tots els incidents relacionats amb agressions de caràcter sexual.

Temps que triguen les víctimes a denunciar

Més enllà de veure com es distribueixen les denúncies al llarg dels dies (Figura 28), sembla més adequat centrar-se en el gràfic de la Figura 29. Es pot concloure que, en general, la major part de les denúncies es realitzen entre 3 i 10 dies posteriors al incident. Tot i que el pic més elevat es troba el primer dia, en el conjunt de dies posteriors n'hi ha més. Si ens fixem llavors en els tipus d'incidents individualment, es pot veure com els robatoris no segueixen aquesta línia. Sinó que la majoria d'ells (38%) es denuncien en les primeres 48 hores, seguit per un 36% denunciat entre

els 3 i 10 dies posteriors. Sembla lògic que les denúncies per robatori es facin el més aviat possible, ja que només és un tràmit policial. Per altra banda, tots els incidents relatius a les agressions sexuals (insults sexuals, coacció i assalt sexual) segueixen el mateix patró: la majoria de denúncies es realitzen entre els 3 i 10 dies posteriors (59%, 63% i 46% respectivament). Entre elles, també es poden distingir entre les que tenen contacte físic i les que són insults. Les agressions amb contacte físic tenen menys denúncies les primeres 48 hores que de 11 a 30 dies després del incident i més de 30 dies. Mentre que amb els insults no passa, hi ha més denúncies les primeres 48h que en les franges de 11 a 30 dies del incident i més de 30 dies. Les agressions amb lesions també segueixen aquesta dinàmica, però els percentatges són menys radicals al llarg del temps que aquests últims. En definitiva, les dades recolzen un concepte lògic, com és que les víctimes que han patit assetjament o agressions de diferents tipus necessiten més temps per a recompondre's i tenir el valor d'explicar-ho i denunciar-ho.

Text

A l'hora d'analitzar els ngrames o paraules clau del text, l'objectiu no era que jo ho analitzés com a tal. Sinó més aviat fer una base o anàlisi preliminar per tantejar les dades extretes, i que els membres del projecte hi treballessin com els semblés oportú. S'han creat els gràfics per a que els analitzin, ja que saben millor què estan buscant i què pot ser interessant. A més a més, no li trobo molt de sentit a analitzar text per part meva perquè el meu nivell d'alemany no és prou bo, i considero que es perdrien matisos de la llengua al intentar treure'n conclusions. Per altra banda, costa analitzar els gràfics amb ngrames i noms comuns perquè la selecció de noms i la neteja de *stopwords* no ha estat prou acurada. De manera que hi ha paraules de les que, des del meu punt de vista, sense un coneixement profund de la llengua, no sembla que se'n pugui deduir massa. Tanmateix, les paraules interessants més comunes es refereixen als llocs dels incidents, els tipus d'aquests, en quines condicions es van donar els incidents o alguns adjectius dels agressors. Si es comparen els gràfics amb els 100 noms més comuns i els 100 noms més comuns a

partir dels 10 dies de denúncia, sembla ser que no hi ha gaire diferència. Els noms més rellevants hi són en ambdues parts.

Sentiment Analysis

Les eines usades per a l'anàlisi de la polaritat serveixen com a primera aproximació al problema. És una disciplina massa complexa per a treure conclusions definitives amb la feina realitzada, i encara se li podria treure molt de suc. Es necessitarien més coneixements de llengua i/o psicologia per a analitzar-ho apropiadament.

Els resultats obtinguts indiquen que no hi ha massa diferències al llarg dels dies (falta de pics pronunciats al gràfic). Malgrat tot, els resultats més negatius es troben les primeres 48 hores i els més positius entre els 3 i 7 dies. Si ens guiem pel tipus d'incident per determinar la polaritat, no tindria molt de sentit, ja que com s'ha vist prèviament, proporcionalment la majoria són robatoris al principi i després vénen les agressions sexuals. Una altra possibilitat podria ser que durant les primeres 48 hores les víctimes tenen l'incident més fresc i es faci servir un to més enfadat degut a això, i més endavant subconscientment estiguin més calmades, tot i en general ser incidents més agressius.

Influència del gènere del/a policia

En relació a la taula de la Figura 44 de la distribució dels tipus de crims entre agents homes i dones, es poden treure un parell de conclusions objectives. La primera és que només un terç del cos policial està format per dones, o almenys aquella part que té contacte amb les víctimes. La segona és que, en línies generals, les dones s'encarreguen un xic més dels incidents que tenen a veure amb agressions sexuals, mentre que els homes ho fan d'altres crims com els robatoris o agressions amb lesions no sexuals. Tanmateix, al no saber com operen ni quines són les proporcions reals agent home/dona a les comissaries de la regió de Colònia, no es pot concloure si aquests resultats són premeditats o bé es dóna de forma natural degut a un esbi-

aixament a l'hora de tractar els crims.

Com s'ha comentat durant l'anàlisi, no es troben diferències significants al comparar els ngrames comuns de les descripcions separades en ambdós gèneres. Al comentar-ho amb els membres del projecte, es va arribar a la conclusió que això podia venir donat per dos motius. L'alemany és un idioma molt formal en sí, també a peu de carrer, i no deixa lloc a trobar diferències en la parla entre ambdós gèneres. Per altra banda, les descripcions dels fets no sempre són cites directes de les víctimes, sinó que sovint són escrites pels oficials com a una transcripció de la denúncia.

Per acabar, analitzant el gràfic de *sentiment analysis* es veu com, novament, no hi ha massa diferència entre el gènere dels oficials. Els pics més destacables, com ara els de dones els dies 13 i 14, són donats perquè es tenen poques mostres en aquests dies i provoca que la precisió al establir-ne la polaritat es vegi afectada. Especialment ja que ambdues tenen únicament agressions sexuals, les quals és normal que tinguin un to més negatiu, però una té polaritat destacablement positiva i la segona negativa. A més a més, es pot destacar la diferència entre oficials homes i dones del dia 9, però comprovant els tipus de crims per a ambdós sexes, no sembla que hi hagi cap correlació que ho expliqui. Ambdós tenen pocs informes en aquell dia i tenen una barreja entre agressions sexuals i robatoris per igual. Pel que fa a la majoria de dies, en especial aquells que tenen més mostres de text per analitzar, s'observa que per a ambdós sexes la línia de polaritat es mostra propera a la general. De manera que no sembla que hi hagi influència en la polaritat de les denúncies es funció de si l'agent de policia és home o dona.

5 Conclusions

Al llarg del treball s'ha anat aconseguint arribar als objectius marcats al principi, com són extreure la informació de les imatges escanejades barrejades en *pdfs* i analitzar-la. S'han creat eines que permeten analitzar el dataset resultant. Crear diferents mapes de calor amb els incidents, filtrar el dataset per diferents paràmetres, analitzar els tipus d'incidents en funció dels dies que triguen les víctimes a denunciar, extreure paraules comunes, observar com paraules clau seleccionades es comporten al llarg dels dies que passen, analitzar la polaritat de les víctimes i buscar si el gènere del policia té algun impacte en la denúncia.

En relació als resultats obtinguts, el fet d'haver fet mineria de dades buscant patrons en diferents apartats, en lloc de tenir una sola hipòtesis i donar-li resposta, dóna lloc a haver aconseguit resultats de diferents aspectes a tractar. Destaquen conclusions com que les víctimes d'agressions i insults sexuals triguen més a denunciar, que els incidents que passen a les afores de la ciutat tenen lloc a les acaballes de la nit i que, en general, les dones policia s'encarreguen més de les denúncies que tenen a veure amb agressions sexuals. Tot i això, des d'un inici l'objectiu no era arribar a resultats determinants, sinó considerar què es podia extreure de les dades. I en aquest aspecte també que crec que s'ha aconseguit una base interessant per a que segueixin treballant.

Finalment, a títol personal, haver treballat en aquest projecte m'ha ensenyat que l'anàlisi de dades és una disciplina molt horitzontal, i que requereix diverses habilitats i flexibilitat a l'hora de treballar. Des dels diferents formats amb els que es poden entregar les dades (poden estar en subòptimes condicions per a treballar, com en aquest cas), a l'ús d'eines estadístiques i la creació de visualitzacions per a poder mostrar els resultats adequadament. A més a més, he tingut la sort de treballar en un cas real, que m'ha aportat experiència i la sensació d'haver estat treballant en quelcom útil al llarg del projecte.

6 Treball Futur

El treball desenvolupat forma part del projecte que tenen en marxa entre el centre de criminologia (KrimZ) de Wiesbaden i la universitat (Johannes Gutenberg Universität). Per això, com que aquest segueix en marxa, s'ha demanat que s'indiqui de quins informes no s'ha pogut extreure el 100% de la informació, per a poder omplir els buits manualment en cas de que es vulgui. I així arribar a conclusions més sòlides i permetre altres futures extensions. S'ha donat un llistat de quines parts dels reports no són extretes correctament i, a més a més, una eina per a poder convertir a text parts concretes d'una imatge seleccionades manualment, amb l'objectiu d'extreure més descripcions sense haver d'escriure-les senceres a mà. Endemés, s'adjunta al codi el manual d'ús per a poder reaprofitar aquest en un futur.

Com a extensió dels anàlisis, es podria estudiar la polaritat dels texts en relació als tipus d'incidents. A primera vista, hauria d'haver-hi alguna correlació, donat que l'impacte sobre la víctima no sembla que hagi de ser el mateix en un robatori que en un assetjament. Tot i això, s'hauria de confirmar la hipòtesis amb les dades. A més, hi ha altres apartats dels reports que no s'han extret i es podrien aprofitar, com el valor dels objectes robats o el contingut dels mateixos. També podria ser interessant per al departament identificar quants informes han estat modificats manualment a posteriori.

Finalment, el que fora una idea per al treball però no es va arribar a desenvolupar, és crear una plataforma que agrupi els anàlisis fets i permeti a l'usuari filtrar el dataset, buscar informes concrets per paraules clau, crear els diagrames, etc. usant una interfície gràfica. Això, junt amb una millora de la visualització de les dades, facilitaria força la feina als investigadors.

Referències

[1] Skymind. Artificial Intelligence Wiki [en línia] [consulta: 31 d'agost de 2018]. Disponible a: <https://skymind.ai/wiki/convolutional-network#intro>

[2] Tensorflow. About Tensorflow. A: *The Tensorflow Homepage* [en línia] [consulta: 30 d'agost del 2018]. Disponible a: <https://www.tensorflow.org/>

[3] Technopedia. Tagged Image File Format (TIFF). [en línia] [consulta: 31 d'agost del 2018]. Disponible a: <https://www.techopedia.com/definition/2093/tagged-image-file-format-tiff>

[4] ImageMagick. [en línia] [consulta: 4 de maig del 2018]. Disponible a: <https://www.imagemagick.org/script/index.php>

[5] Rouse, Margaret. *OCR (optical character recognition)*. A: TechTarget. Setembre de 2005. [en línia] [consulta: 31 d'agost del 2018]. Disponible a: <https://searchcontentmanagement.techtarget.com/definition/OCR-optical-character-recognition>

[6] Github Tesseract-ocr. Wiki tesseract: Improve Quality. [en línia] [consulta: 21 de maig del 2018]. Disponible a <https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality>

[7] Image Processing Learning Resources. Erosion. [en línia] [consulta: 18 de juny del 2018]. Disponible a: <https://homepages.inf.ed.ac.uk/rbf/HIPR2/erode.htm>

[8] OpenCV. Cool theory, Morphological Operations. [en línia] [consulta: 18 de juny del 2018]. Disponible a: <https://docs.opencv.org/2.4/doc/tutorials/imgproc/erosion_dilatation/erosion_dilatation.html>

[9] OpenCV. Template Matching. [en línia] [consulta: 14 de maig del 2018]. Disponible a: <https://docs.opencv.org/2.4/doc/tutorials/imgproc/histograms/template_matching/template_matching.html>

[10] OpenCV. Smoothing Images, Median Filter. [en línia] [consulta: 22 de maig del 2018]. Disponible a: <https://docs.opencv.org/2.4/doc/tutorials/imgproc/gausian_median_blur_bilateral_filter/gausian_median_blur_bilateral_filter.html>

[11] Zeokat. *¿Qué son las stop words o palabras vacias?*. A: VozIdea. 7 de julio de 2014. [en línia] [consulta: 31 d'agost del 2018]. Disponible a: <http://www.vozidea.com/que-son-las-stop-words-o-palabras-vacias>

[12] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. [en línia] [consulta: 21 de juny del 2018]. Disponible a: <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>

[13] NLTK. Pàgina principal. [en línia] [consulta: 21 de juny del 2018]. Disponible a: <http://www.nltk.org/>

[14] Documentació Geopy. Pàgina principal. [en línia] [consulta: 12 de juny del 2018]. Disponible a: <https://geopy.readthedocs.io/en/stable/#>

[15] Barbr, Jordan. *Latent Dirichlet Allocation (LDA) with Python*. [en línia] [consulta: 22 de juny del 2018]. Disponible a: <https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html>

[16] Lexalytics. Sentiment Analysis. [en línia] [consulta: 23 d'agost del 2018]. Disponible a: <https://www.lexalytics.com/technology/sentiment>

[17] Tatman, Rachael. *German Sentiment Analysis Toolkit*. [en línia] [consulta: 14 de juliol del 2018]. Disponible a: <https://www.kaggle.com/rtatman/german-sentiment-analysis-toolkit#SentiWS_v1.8c_Positive.txt>

[18] SePL (Sentiment Phrase List). [en línia] [consulta: 18 de juliol del 2018]. Disponible a: <http://www.opinion-mining.org/index.php/ger/SePL-Sentiment-Phrase-List>